# 结构稀疏模型

刘建伟 崔立鹏 罗雄麟
(中国石油大学(北京)自动化系 北京 102249)

**摘 要** 由于生物信息学、心理学诊断、计算语言与语音学、计算机视觉、门户网站、电子商务、移动互联网、物联网中处理高维和超高维数据的需求不断涌现,迫切需要研究具有变量选择和特征降维功能的回归和分类模型,所以以 Lasso、自适应 Lasso 和 elastic net 等为代表的稀疏模型近年来在机器学习领域中非常流行.然而,这些稀疏模型没有考虑变量中存在的组结构、重叠组结构、双层稀疏结构、多层稀疏结构、树结构和图结构等结构化信息.结构稀疏模型考虑了这些结构先验信息,改善了模型对特征选择的结果和稀疏模型在相应结构稀疏化数据背景下的统计特性.结构稀疏化模型是当前稀疏学习领域的研究方向,近几年来涌现出很多研究成果,文中对主流的结构稀疏模型,如组结构稀疏模型、结构稀疏字典学习、双层结构稀疏模型、树结构稀疏模型和图结构稀疏模型进行了总结,对结构稀疏模型目标函数中包含非可微、非凸和不可分离变量的结构稀疏模型目标函数近似转换为可微、凸和可分离变量的近似目标函数的技术如控制-受控不等式(Majority-Minority,MM),Nesterov 双目标函数近似方法,一阶泰勒展开和二阶泰勒展开技术,对求解结构稀疏化模型近似目标函数的优化算法如最小角回归算法、组最小角回归算法(Group Least Angle Regression,Group LARS)、块坐标下降算法(block coordinate descent algorithm)、分块坐标梯度下降算法(block coordinate gradient descent algorithm)、局部坐标下降算法(local coordinate descent algorithm)、谱投影梯度法(Spectral Projected Gradient algorithm)、主动集算法(active set algrithm)和交替方向乘子算法(Alternating Direction Method of Multipliers,ADMM)进行了比较分析,并且对结构稀疏模型未来的研究方向进行了探讨.

**关键词** 稀疏化模型;结构稀疏化模型;组结构稀疏模型;多层稀疏结构模型;树结构稀疏化模型;图结构稀疏化模型;结构稀疏字典;结构稀疏码;人工智能
**中图法分类号** TP18 **DOI 号** 10.11897/SP.J.1016.2017.01309

# Structured Sparse Models

LIU Jian-Wei CUI Li-Peng LUO Xiong-Lin
(*Department of Automation, China University of Petroleum, Beijing* 102249)

**Abstract** As continuing to emerge demand of high dimensional and ultra-high dimensional regression and classification in bioinformatics, psychology diagnosis, computational linguistics and phonetics, computer vision, the Portal site, e-commerce, mobile Internet, and Internet of Things, there is an urgent need to study high dimensional and ultra-high dimensional variable selection and feature dimension reduction in regression and classification model. Thus the sparse models have been quite popular in recent years, such as the Lasso, adaptive Lasso and the elastic net. However, these sparse models ignore the structural information of the variables, such as the group structure sparsity, overlapping group structure sparsity, bi-level sparse structure, Multi-layer Sparse structure, tree structure sparsity and graph structure sparsity. The structured sparse models that consider this structural prior information can improve the statistic properties of the sparse models when facing with the corresponding structure sparse datasets. The structured

收稿日期:2016-01-25;在线出版日期:2016-09-27.**刘建伟**,男,1966 年生,博士,副研究员,主要研究方向为机器学习、智能信息处理、复杂系统的分析、预测与控制. E-mail: liujw@cup.edu.cn.**崔立鹏**,男,1990 年生,硕士研究生,主要研究方向为机器学习.**罗雄麟**,男,1963 年生,博士,教授,主要研究领域为智能控制和复杂系统的分析、预测与控制.

sparse models are the hot research direction of the sparse model learning and many research findings appear in recent years. This paper gives a systematic survey of mainstream of structured sparsity model, such as group structure sparse model, structure sparse dictionary learning, bi-level structure sparse model, and tree structure sparse model and graphical structure sparse model. As objective function of structure sparse model contains non-differential, non-convex and non-separable variable, objective function of structure sparse model first needs to be approximately transform into differentiable, convex and separable variable ones. The main approximate transformation methods are summarized, including majority-minority inequality, approximate method of Nesterov's double objective function, first order Taylor expansion and second order Taylor expansion. Optimization algorithms solving approximate objective function of structure sparse model are carried out a detailed comparative analysis on the conception, the features and performance, which involves minimum angle regression, group Least angle regression, block coordinate descent algorithm, block coordinate gradient descent algorithm, local coordinate descent algorithm, spectrum projection gradient method, active set algorithm and alternating direction method of multipliers, some future research directions are discussed in the final section.

**Keywords** sparsity model; structured sparsity model; group structure sparsity model; multi-layer Sparse structure model; tree structure sparse model; graph structure sparse model; structured sparse dictionary learning; structured sparse coding; artificial intelligence

## 1 引 言

在统计机器学习研究中,需要学习高维数据的模型,样本的维数高达数十万维,如果强行在高维数据上建立高维模型,建立的模型就会失去模型的解释性,找不到与输出变量最相关的输入影响因素.另外根据人类大脑认知、学习和推理的过程可知,高维数据总是在低维空间中进行学习和推理,信息论、编码理论和压缩传感理论也说明,高维数据完全可以从低维数据重构出来,而且能够在概率意义上统计得知担保重构误差很小.

另一个原因是统计学中发现模型降维,即只选择少量的预测变量比精确地求解包含全部预测变量的最小二乘和极大似然估计问题,能够解决模型过拟合问题.但是难点在于选择几个预测变量,选择哪几个变量,有的学者提出用 Cp、AIC 和 BIC 等准则确定所选变量的个数.

套索模型(Least absolute shrinkage and selection operator, Lasso)[1]能够同时实现变量选择和模型参数估计. Tibshirani 提出的套索为

$$\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\cdot\|\boldsymbol{\beta}\|_1 \tag{1}$$

其中 $\lambda\geqslant 0$, $\|\cdot\|_2$ 表示 $L_2$ 范数, $\|\boldsymbol{\beta}\|_1=\sum_{p=1}^{P}|\beta_p|$ 表示向量 $\boldsymbol{\beta}$ 的 $L_1$ 范数, $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$, $\boldsymbol{y}\in\boldsymbol{R}^N$, $\boldsymbol{\beta}\in\boldsymbol{R}^P$ 叫做模型向量, $N$ 为样本个数, $P$ 为变量个数.套索的解为满足以下条件的 $\hat{\beta}_i$

$$\begin{aligned}&\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\lambda\,\mathrm{sign}(\hat{\beta}_i)\quad \text{if}\quad \hat{\beta}_i\neq 0\\&|\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})|\leqslant\lambda\quad \text{if}\quad \hat{\beta}_i=0\end{aligned} \tag{2}$$

这里 $X_i$ 为矩阵 $\boldsymbol{X}$ 第 $i$ 列, sign(·)为符号函数.当 $\hat{\beta}_i>0$ 时, $\mathrm{sign}(\hat{\beta}_i)=1$; $\hat{\beta}_i<0$ 时, $\mathrm{sign}(\hat{\beta}_i)=-1$;其他情况下, $\mathrm{sign}(\hat{\beta}_i)=0$.

满足不等式 $|\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})|\leqslant\lambda$ 的模型向量的分量的解为 0, $\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ 落在区间 $[-\lambda,\lambda]$ 内时,模型向量 $\boldsymbol{\beta}$ 的分量为 0.

套索模型的模型向量的解随着 $\lambda$ 的变化而变化,有无数个解.这些解随着 $\lambda$ 而变化,画出一条曲线,为套索模型的解路径.可以证明套索的解路径是一条分段线性曲线,假定 $\lambda_1<\lambda_2$, $\lambda_1$ 与 $\lambda_2$ 足够靠近,只要 $\hat{\boldsymbol{\beta}}$ 的符号不变,则解 $\hat{\boldsymbol{\beta}}(\lambda_1)$ 和 $\hat{\boldsymbol{\beta}}(\lambda_2)$ 不改变,即两者相等.也就是如果 $0\leqslant a\leqslant 1$,则对所有的 $\lambda=a\lambda_1+(1-a)\lambda_2$ 和 $\hat{\boldsymbol{\beta}}=a\hat{\boldsymbol{\beta}}(\lambda_1)+(1-a)\hat{\boldsymbol{\beta}}(\lambda_2)$, $\hat{\boldsymbol{\beta}}(\lambda)=\hat{\boldsymbol{\beta}}$,也就是只要模型向量的解的符号不变,在 $\lambda_1$ 与 $\lambda_2$ 之间的解路径是一条直线.这样,可以通过求解解路径,比较所有的解,使得误差最小的解为最终的套索解.

套索只能实现普通的非结构稀疏化效果,其示意图如图 1 所示,从 7 个变量中选择出的变量不具有任

何结构化的形式，然而有时样本的多个变量之间本身具有某种结构，来看一个实例，Hosmer 等人[2]收集的新生儿体重数据集(Birthweight Dataset)，见表 1. 该数据集包含 189 个新生儿的体重以及可能与新生儿体重有关的一些解释变量：母体年龄、母体体重、种族、吸烟史、早产史、高血压史、子宫刺激性史以及怀孕期间的物理检查次数，其中母体年龄和母体体重为连续变量，而其余 6 个解释变量均为取离散值的分类变量. 对于取离散值的分类变量，其所对应的多个水平值可被视为多个特征变量，这些属于同一分类变量的特征变量自然形成组变量，这就是一种典型的组结构模型的情形. 又例如使用 fMRI 数据进行精神疾病分析，图 1 中，输出 $\boldsymbol{Y}$ 为要预测的精神状态、人类行为或执行特定任务的外部应激反应，矩阵 $\boldsymbol{X}$ 为 fMRI 数据，矩阵 $\boldsymbol{X}$ 的列由 30 万个三维像素特征变量组成，矩阵 $\boldsymbol{X}$ 的行有 1000 个样例，大脑的不同脑功能区域，自然地把 30 万个三维像素特征变量分成多个组，不同脑功能区域对应的模型向量 $\boldsymbol{\beta}$ 分量自然地分成组. 此外，还有很多组稀疏化情形，在基因微阵列数据分析中起同一生物功能的多个基因序列看做一个组；在基因关联研究中，某基因的全部基因标记可以被看作一个组；在多因子方差分析中，因子可能同时具有若干个水平，因而需要用一组哑变量来表示这些水平，因子对应的这些哑变量可以被看做一个组；在信号处理邻域，小波被广泛地用作信号表示的基函数，实验发现信号小波表示的能量集中于很少几个系数上，而且选择小波基个数时，由于小波系数存在空间上的近邻关系，自然地形成树结构；基于主题模型进行图像重构时，属于同一主题的图像区域被自然地分为一组；对于有类标签的数据，使用二值指示变量编码表示分类变量所属的类，这样自然地把变量被分为组；多任务学习中，使用同样模型参数的输入变量预测多个目标变量，具有同样模型参数的输入变量自然地属于同一组. Tibshirani 提出的套索模型没有利用上述先验结构信息，得不到结构稀疏模型，得到的模型没有考虑变量同时被选中或同时不被选中的情况，这种结构关系，如图 2 所示.

**表 1　新生儿体重数据集中的某些变量**

| 种族 | | 吸烟史 | | 早产史 | | 高血压史 | | 子宫刺激性史 | | 怀孕期间的物理检查次数 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 白人 | 黑人 | 吸烟 | 不吸烟 | 一次 | 两次 | 有 | 无 | 有 | 无 | 一次 | 二次 | 三次 |

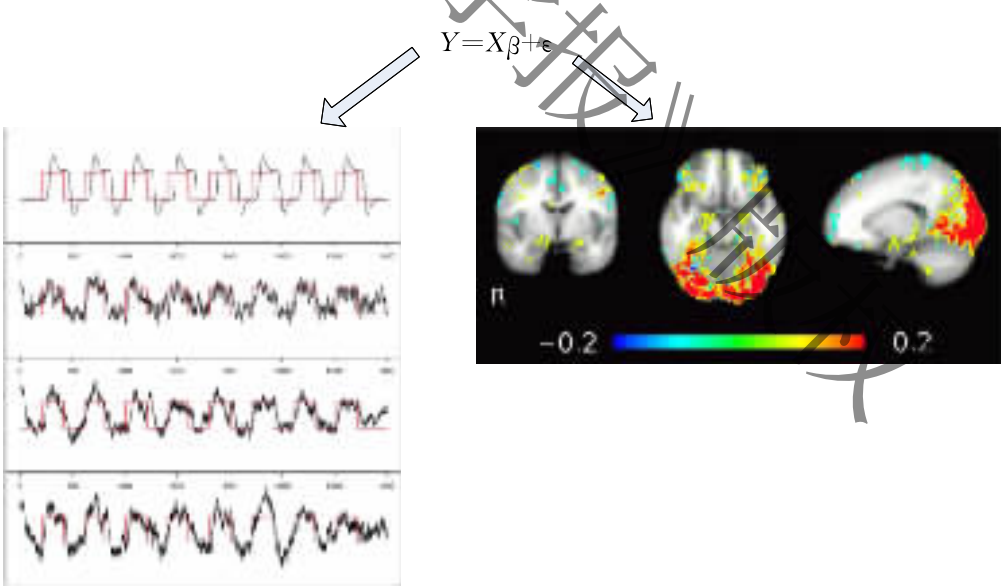



图 1　fMRI 数据集上神经状态预测模型

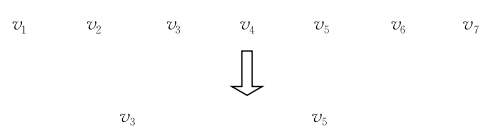



图 2　套索的稀疏化效果

图 2 中令 $\boldsymbol{\beta}=v=(v_1,\cdots,v_P)^{\mathrm{T}}\in\boldsymbol{R}^P$，$v_3$、$v_6$ 和 $v_{11}$ 为选中的变量，其余未选中的变量对应的模型分量 $\beta_p$ 均为 0. Yuan 等人[3]充分利用这种组结构作为先验信息，将变量事先进行分组，再与 Tibshirani 提出的套索模型的 $L_1$ 范数罚方法相结合，提出了组套索(Group Lasso)，组套索在变量组的水平上同时实现了稀疏变量组选择和模型参数估计. 显然，Tibshirani 提出的套索和 Yuan 等人提出的组套索之间的区别在于模型选择的效果不同，套索模型导

致的模型稀疏性是变量水平上的，未选中的变量对应的模型分量 $\beta_p$ 均为 0；而组套索导致的模型稀疏性是变量组水平上的，这些未选中的组中的变量对应的模型向量 $\boldsymbol{\beta}_j$ 均为 0，即变量组选择. 组套索的提出引出了利用结构稀疏化思想进行稀疏变量选择的新思路：结构稀疏模型. 结构稀疏模型的目标函数的一般形式为[4-8]

$$\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P} \Xi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})=\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})+\lambda\cdot\Omega(\boldsymbol{\beta}) \quad (3)$$

其中 $\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$ 为最小二乘、铰链或逻辑斯蒂损失函数，$\Omega(\boldsymbol{\beta})$ 为产生各种结构稀疏模型的罚函数，$\lambda\geq 0$，$\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$，$\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_P)^{\mathrm{T}}\in\boldsymbol{R}^P$. 回归时，$\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$ 为最小二乘损失函数，$\boldsymbol{y}=(y_1,y_2,\cdots,y_N)^{\mathrm{T}}\in\boldsymbol{R}^N$，分类时 $\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$ 为铰链或逻辑斯蒂损失函数，$\boldsymbol{y}\in\{-1,+1\}$. 式(3)还可以改为以下有约束的等价形式：

$$\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P} \Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y}) \quad (4)$$
$$\text{s.t.} \quad \Omega(\boldsymbol{\beta})<m^*$$

这里 $m^*$ 为决定解路径的阈值. 构造各种诱导组稀疏模型的罚函数 $\Omega(\boldsymbol{\beta})$ 是结构稀疏模型的主要任务.

不同的结构稀疏化模型所实现的稀疏化效果往往不同，因此本文按照稀疏模型所实现的稀疏化效果对其进行分类，即按照其实现的稀疏化结构将其分为组结构稀疏模型、结构稀疏字典学习、双层结构稀疏模型、树结构稀疏模型和图结构稀疏模型这几个部分来进行讨论，其具体分类如图 3 所示[9-18].

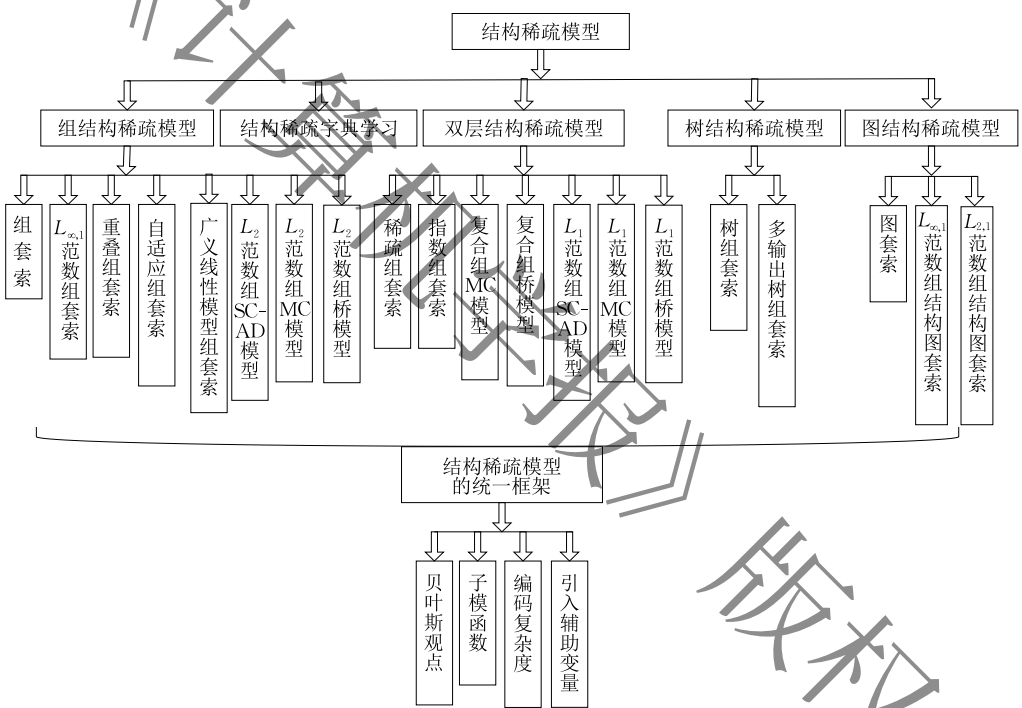



图 3 结构稀疏模型分类图

## 2 组结构稀疏模型

### 2.1 组套索模型

组模型典型的例子如用二值指示变量表示分类输入变量所属的类，同一类输入变量，有一个组子模型向量；对于非参数加回归模型，组模型子向量 $\boldsymbol{\beta}_{g_j}$ 为第 $j$ 个输入变量的按第 $j$ 个加函数的基展开系数；再比如同一输入变量使用多组模型向量预测多个输出变量的多任务学习. 又比如在多变量多水平因子分析模型中，每个水平因式上的模型向量不同，假定输入 $\boldsymbol{X}$ 为四水平因子 $g_1g_2g_3g_4$ 模型 $G$，则线性模型的均值为

$$E(\boldsymbol{Y}|\boldsymbol{X},G)=\boldsymbol{X}\boldsymbol{\beta}+\sum_{k=1}^{4}\beta_{0k}I_{g_k}(G)$$

引入虚拟(哑)变量 $\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4)$，令 $\pi_k=I_{g_k}(G)$，$\beta_0=(\beta_{01},\beta_{02},\beta_{03},\beta_{04})$. 上述模型可表示为 $E(\boldsymbol{Y}|\boldsymbol{X},G)=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\pi}^{\mathrm{T}}\beta_0$，$\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3,\pi_4)$ 为单个水平 $g_i$ 上的组模型子向量.

假定模型为

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon \quad (5)$$

其中 $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$，$\boldsymbol{\beta}\in\boldsymbol{R}^P$，$\boldsymbol{y}\in\boldsymbol{R}^N$，$\varepsilon\in\boldsymbol{R}^N$ 且 $\varepsilon\sim N(0,\sigma^2\boldsymbol{I})$. 这里，$N$ 为样例个数，$n\in\{1,2,\cdots,N\}$ 表示样例或观察次数的索引. $P$ 为变量个数，用 $p\in\{1,\cdots,$

$P\}$表示变量的索引. 设模型向量的 $P$ 个分量被分为 $J$ 个组 $G=\{g_j\mid j=1,2,\cdots,J\}$，$\boldsymbol{\beta}=\bigcup_{j=1}^{J}g_j$，对于 $i\neq j$，$g_i\cap g_j=\varnothing$. $d_j$ 表示第 $j$ 个组中变量的个数，$\boldsymbol{\beta}\in\boldsymbol{R}^P$ 叫做模型向量，$\boldsymbol{\beta}_j\in\boldsymbol{R}^{d_j}$ 叫做第 $j$ 个组 $\boldsymbol{X}_j$ 对应的子模型向量，组套索模型的目标函数为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_{K_j}\tag{6}$$

其中，$\|\boldsymbol{\beta}_j\|_{K_j}=\sqrt{\boldsymbol{\beta}_j^{\mathrm{T}}K_j\boldsymbol{\beta}_j}$，$\lambda\geq0$，当取 $K_j=n_j\boldsymbol{I}_{n_j}$，$\boldsymbol{I}_{n_j}\in\{0.1\}^{n_j\times n_j}$ 为单位矩阵时，得到

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2,$$

这里，$\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$ 为块 $L_1/L_2$ 罚，随着 $\lambda$ 的改变，子模型向量分量要么全部为 0，要么所有的子模型向量分量均不为 0. 其中，$\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$ 为 $L_{2,1}$ 范数，简写为 $\|\boldsymbol{\beta}\|_{2,1}$. 当第 $j$ 个组中变量的个数 $d_j$ 为 1 时，$\|\boldsymbol{\beta}_j\|_2=|\boldsymbol{\beta}_j|$，此时，每组内只有一个变量，式(6)变为套索模型.

组套索的变量选择效果如图 4 所示，其中每个椭圆形中的变量对应一个组.

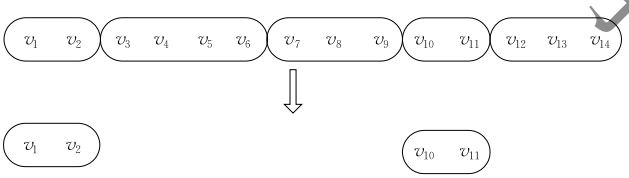



图 4　组套索的稀疏化效果

## 2.2　$L_{\infty,1}$ 组套索

$L_{\infty,1}$ 组套索将组套索中的罚函数由 $L_1$ 范数和 $L_2$ 范数的组合变为 $L_1$ 范数和 $L_\infty$ 范数的组合：先对组 $j$ 对应的子模型向量 $\boldsymbol{\beta}_j$ 执行 $L_\infty$ 范数运算，再对由子模型向量 $\boldsymbol{\beta}_j$（其中 $j\in\{1,\cdots,J\}$）的 $L_2$ 范数值组成的向量求 $L_1$ 范数，便得到 $L_{\infty,1}$ 范数罚 $\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_\infty$. 已知式(5)中的线性回归模型，则 $L_{\infty,1}$ 组套索具有如下的形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_\infty\tag{7}$$

其中 $\lambda\geq0$，第 $j\in\{1,\cdots,J\}$ 个组子模型向量为 $\boldsymbol{\beta}_j$，第 $j$ 个组中变量的势为 $d_j$. 实际上组套索和 $L_{\infty,1}$ 组套索可以归结为以下的 $L_{q,1}$ 组套索[19,20]：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_q\tag{8}$$

这里，$1\leq q\leq\infty$，$\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_q$ 为 $L_{q,1}$ 范数. 显然，在式(8)中，当 $q=1$ 时对应 Tibshirani 的套索模型，当 $q=2$ 时对应组套索，当 $q=\infty$ 时对应 $L_{\infty,1}$ 组套索. Vogt 等人[21]对 $L_{\infty,1}$ 组套索和组套索进行了仿真比较，指出组套索在预测准确性、计算复杂性和解的可解释性方面都优于 $L_{\infty,1}$ 组套索.

## 2.3　重叠组套索

组套索在实际应用中具有局限性，例如在微阵列基因表达数据分析中，由于基因以多路径(multiple pathways)的方式参与模型的构造过程，因此一个基因可能同时属于多个组，即组与组之间包含的变量存在重叠. 非重叠组稀疏模型没有考虑不同的组子模型向量可能包含相同的某些变量，这些变量组合情况被非重叠组稀疏模型排除在外了. 假设 $P$ 个变量被分为 $J$ 个组 $g_1,\cdots,g_J$，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示组的索引集，用 $\boldsymbol{\beta}_{g_j}$ 表示组 $g_j$ 对应的子模型向量，$G=\{g_j\mid j=1,\cdots,J\}$ 表示全部组的索引集的集合，且 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$. 注意上述定义中 $g_j\subseteq\{1,\cdots,P\}$ 允许不同组之间的变量出现重叠，引入了“重叠”结构这种先验信息的方法. 已知式(5)中的线性回归模型，重叠组套索求解目标函数为[21-24]

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2\tag{9}$$

这里，$w_{g_j}$ 为组 $g_j$ 固定权值，$g_j\subseteq\{1,\cdots,P\}$. Obozinski 等人指出[24]，当组具有重叠结构时，若利用式(4)即组之间变量不具有重叠结构的组套索进行组内的变量选择的话，因为未选中组 $\{v_1,v_2,v_3,v_4\}$，故在丢弃组 $\{v_1,v_2,v_3,v_4\}$ 的时候将变量 $v_3$ 和 $v_4$ 都丢弃了. 而组 $\{v_3,v_4,v_5,v_6\}$ 被选中，但由于丢弃了变量 $v_3$ 和 $v_4$，所以最终的选择效果中将不会含有变量 $v_3$ 和 $v_4$，如图 5 所示. 导致这种情况的原因是没有考虑组 $\{v_1,v_2,v_3,v_4\}$ 和组 $\{v_1,v_2,v_3,v_4\}$ 之间存在重叠的变量 $v_3$ 和 $v_4$ 这种先验信息. 重叠组套索的变量选择效果如图 6 所示，组 $\{v_1,v_2,v_3,v_4\}$ 被完整地选择出来了. 另外，Jenatton 等人[25]均指出重叠组套索在满足一定条件时具有变量选择一致性，并给出了满足变量选择一致性所需要的假设条件. Percival 等人[26-28]给出了输出预测误差 $\|\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|$ 和估计值的 $L_{2,1}$ 范数误差 $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{2,1}$ 的上界，同时还指出在变量个数一定的前提下，划分的组数目越多则估计值的误差上界就越大. 因此其建议在使用重叠组套索时要考虑组的划分情况，宜在划分的组的个数不太大的情况下使用重叠组套索.

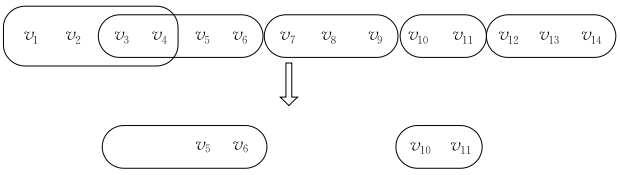



图 5 组套索在数据具有重叠组结构时的稀疏化效果

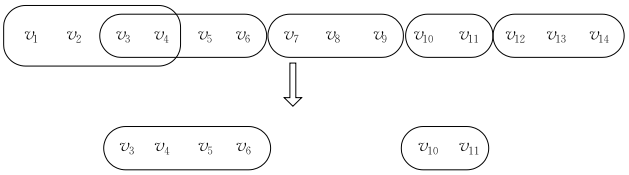



图 6 重叠组套索在数据具有重叠组结构时的稀疏化效果

### 2.4 广义线性模型组套索

可以通过定义 $\mu=\boldsymbol{X\beta},\Sigma=\boldsymbol{I},\boldsymbol{y}\sim N(\mu,\Sigma)$，求解 $\boldsymbol{y}$ 的负对数似然函数[29]

$$-\log p(\boldsymbol{y}|\boldsymbol{X\beta})=T(\boldsymbol{y})+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2 \quad (10)$$

即可得到线性模型最小二乘目标函数 $\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2\leq\varepsilon$. 这是假定线性模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\varepsilon$ 的噪声 $\varepsilon$ 为高斯分布的例子. 假定模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\varepsilon$ 中的噪声 $\varepsilon$ 为指数分布簇 $p(\boldsymbol{y})=p_0(\boldsymbol{y})\mathrm{e}^{\theta^{\mathrm{T}}T(\boldsymbol{y})-\psi(\theta)}$，如伯努利分布、多项分布、指数分布、伽马分布、卡方分布、贝塔分布、韦伯分布(Weibull)、狄氏分布和泊松分布，令

$$\mu(\theta)=E(\boldsymbol{y})=f^{-1}(\boldsymbol{X\beta})=f^{-1}(\theta)=\nabla\psi(\theta),$$

这里 $f(\cdot)$为线性或非线性连接函数，$\theta$ 为自然参数，$T(\boldsymbol{y})$为 $\boldsymbol{y}$ 的充分统计量，$\psi(\theta)=\log\int p_0(\boldsymbol{y})\mathrm{e}^{\theta^{\mathrm{T}}T(\boldsymbol{y})}\mathrm{d}\boldsymbol{y}$ 为严凸可微对数配分函数，$\nabla\psi(\theta)$为梯度函数，$p_0(\boldsymbol{y})$为基测度，求解负对数似然函数 $-\log p(\boldsymbol{y}|\boldsymbol{X\beta})$即得到相应的各种广义线性模型.

由于 Bregman 离差

$$d_\varphi(\boldsymbol{y},\mu(\boldsymbol{X\beta}))=\varphi(\boldsymbol{y})-\varphi(\mu(\boldsymbol{X\beta}))-(\boldsymbol{y}-\mu(\boldsymbol{X\beta}))^{\mathrm{T}}\nabla\varphi(\boldsymbol{y}) \quad (11)$$

和指数分布簇的等价表示形式

$$p_{\psi,\theta}(\boldsymbol{y})=\mathrm{e}^{(-d_\phi(\boldsymbol{y},\mu(\boldsymbol{X\beta})))f_\phi(\boldsymbol{y})}$$

有一一对应关系，这里 $\phi(\cdot)$为 $\psi(\cdot)$的 Legendre 对偶函数，所以广义线性模型组套索的目标函数为

$$\hat{\boldsymbol{\beta}}(\lambda)=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p}\frac{1}{2}\sum_{i=1}^{N}d_\phi(\boldsymbol{y}_i,\mu(\boldsymbol{X}_i\boldsymbol{\beta}))+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \quad (12)$$

当取连接函数为

$$f(\mu)=\log\frac{\mu}{1-\mu},$$

$$f^{-1}(\theta)=\frac{1}{1+\exp^{-\theta}}$$

时，得到逻辑斯蒂模型

$$E(\boldsymbol{y}|\boldsymbol{X\beta})=\frac{1}{1+\exp^{-\boldsymbol{X\beta}}}.$$

逻辑斯蒂模型组套索模型为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p}\Xi(\boldsymbol{\beta})=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p}\Big\{-\Phi(\boldsymbol{\beta})+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2\Big\},$$

$$\Phi(\boldsymbol{\beta})=\sum_{n=1}^{N}\boldsymbol{y}_n\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta}-\log[1+\exp(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta})].$$

### 2.5 $L_2$ 范数组 SCAD 模型

$L_2$ 范数组 SCAD(Smoothly Clipped Absolute Deviation)罚模型为[30]

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+\sum_{j=1}^{J}\phi_\lambda(\|\boldsymbol{\beta}_j\|_2) \quad (13)$$

其中

$$\phi_\lambda(\theta)=\begin{cases}\lambda\theta, & 0<\theta\leq\lambda\\ -\dfrac{(\theta^2-2\gamma\lambda\theta+\lambda^2)}{2(\gamma-1)}, & \lambda<\theta<\gamma\lambda\\ \dfrac{(\gamma+1)\lambda^2}{2}, & \theta>\gamma\lambda\end{cases} \quad (14)$$

这里 $\gamma>2,\lambda\geq0$. 当 $\gamma\to\infty$时，式(13)等于 $\phi_\lambda(\theta)=\lambda\theta$，因此 $\gamma\to\infty$时，$L_2$ 范数组 SCAD 罚模型变为组套索. 式(14)中的 SCAD 罚[31]是一种非凸的罚函数，其对模型向量分量的惩罚程度随着分量大小的增大而逐渐减小，因此避免了对目标变量回归系数的较大惩罚，构造的模型往往具有变量选择一致性. 对于 $L_2$ 范数组 SCAD 罚模型来说，由于在组水平上使用了非凸的 SCAD 罚，因此 $L_2$ 范数组 SCAD 罚模型具有变量组选择一致性.

### 2.6 $L_2$ 范数组桥模型

已知式(5)中的线性回归模型，$L_2$ 范数组桥模型[32]具有如下的形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+\lambda\sum_{j=1}^{J}c_j\|\boldsymbol{\beta}_j\|_1^\gamma \quad (15)$$

其中 $\lambda\geq0,0<\gamma<1,\boldsymbol{\beta}_j$为第 $j\in\{1,\cdots,J\}$个组的子模型向量，$c_j=d_j^{1-\gamma}$. 当 $d_j=1$ 时，组桥模型即为桥模型[33]，而实际上组桥模型的提出也正是受到了桥模型的启发，实际上是将桥模型推广到组变量选择情形下得到的. 与 SCAD 罚类似，桥罚函数也是一种随着模型向量分量增大而逐渐减小惩罚程度的非凸罚，其构造出的模型也往往具有变量选择一致性. 对于 $L_2$ 范数组桥模型来说，由于在组水平上使用了非凸的 SCAD 罚，因此 $L_2$ 范数组桥模型具有变量组选择一致性.

### 2.7 $L_2$ 范数组 MC 模型

MC(Minimax Concave penalties)罚函数为

$$\sum_{j=1}^{P} P_{\lambda,\gamma}(\boldsymbol{\beta}_j),$$

这里，$P_{\lambda,\gamma}(\boldsymbol{\beta}_j)=\int_0^{|\boldsymbol{\beta}_j|}\left(1-\frac{x}{\lambda\gamma}\right)_+ \mathrm{d}x$.

使用 MC 罚函数，$L_2$ 范数组 MC 罚（Minimax Concave penalties）模型为[34]

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2) \quad (16)$$

这里 $\lambda\geq 0$，$\gamma>1$，

$$\varphi_{\lambda,\gamma}(\theta)=\begin{cases}\lambda\theta-\dfrac{\theta^2}{2\gamma}, & 0<\theta<\gamma\lambda\\ \dfrac{1}{2}\gamma\lambda^2, & \theta>\gamma\lambda\end{cases} \quad (17)$$

由式(17)，令 $\gamma\to\infty$、$\phi_\lambda(\theta)=\lambda\theta$，由此可以得到结论：$L_2$ 范数组 MC 模型在 $\gamma\to\infty$ 时近似为组套索. 式(17)中的 MC 罚[35]与 SCAD 罚类似，也是一种非凸的罚函数，其对模型向量分量的压缩程度随着分量大小增大而逐渐较小，构造的模型也往往具有变量选择一致性. 对于 $L_2$ 范数组 MC 罚模型来说，由于在组水平上使用了非凸的 MC 罚，因此 $L_2$ 范数组 MC 罚模型具有变量组选择一致性.

## 3　多层结构稀疏模型

### 3.1　双层稀疏模型

所谓双层结构稀疏模型指的是既能够实现变量组稀疏又能够实现变量组内变量选择的稀疏模型，即同时实现组稀疏性和组内变量的稀疏性，其变量选择的示意图如图 7 所示，变量选择的结果为从 5 个变量组中选择出了 2 个变量组，与此同时被选择出的变量组中也并非包含原变量组中的全部变量，换句话说既具有组稀疏性又具有组内稀疏性. 例如，在文献[36]的表 3 中，不仅期望探究哪些组变量对体质指数有显著影响，而且期望探究哪些组变量中的哪个变量对体质指数具有显著影响；再比如，在遗传关联研究中，每个基因上都有多个变异点，这些位于同一个基因上的变异点自然应该被视为位于同一个分组，因而希望在识别出与疾病发生相关的产生变异的基因的同时也识别出基因内部与疾病发生相关的变异点[37]. 双层结构稀疏模型大体可被分为两类：罚函数为加和形式的稀疏组套索和 $L_1$ 范数组 SCAD 模型等一类由层次罚函数构成的双层结构稀疏模型.

3.1.1　稀疏组套索

套索导致的模型稀疏性是变量水平上的稀疏性，能够实现组内的变量选择；组套索导致的模型稀疏性是组水平上的稀疏性，能够实现组内的变量选择. 然而很多时候我们既希望将重要的变量组选择出来，又希望将变量组中重要的变量选择出来，组套索显然不能满足这种要求. 如果将 $L_1$ 范数罚添加到组套索的目标函数中，那么既可以实现变量组水平上的模型稀疏性，又可以实现变量水平上的模型稀疏性，即同时实现变量组选择和组内的变量选择，我们将这种组套索叫做稀疏组套索. 已知式(5)中的线性回归模型，令 $j\in\{1,\cdots,J\}$ 表示变量组的索引，$p\in\{1,\cdots,P\}$ 表示变量的索引，则稀疏组套索问题（Sparse Group Lasso）[38]为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2+\lambda_2\|\boldsymbol{\beta}\|_1 \quad (18)$$

其中 $\lambda_1\geq 0$、$\lambda_2\geq 0$. $\boldsymbol{\beta}_j$ 表示第 $j$ 个组对应的子模型向量，罚项 $\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$ 约束组向量个数和组变量内各变量的下标索引值；罚项 $\lambda_2\|\boldsymbol{\beta}\|_1$ 约束组子模型向量的分量的大小. 稀疏组套索可被看作是两层的树组套索，$\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$ 看作是根节点约束，$\lambda_2\|\boldsymbol{\beta}\|_1$ 为叶子节点约束. Chatterjee 等人[39]指出稀疏组套索在满足一定条件时具有参数估计一致性，并给出了满足参数估计一致性所需要的假设条件.

3.1.2　由层次罚构成的一类双层稀疏模型

图 7 所示为双层结构稀疏模型的稀疏化效果. Huang 等人[19]构造了一种层次罚函数

$$\sum_{j=1}^{J} f_{\mathrm{out}}\left\{\sum_{p=1}^{d_j} f_{\mathrm{in}}(\beta_{jp})\right\} \quad (19)$$

其中 $d_j$ 为第 $j$ 个变量组所含的变量数，并且要求内部的罚函数 $f_{\mathrm{in}}$ 在 $[0,+\infty)$ 上为凹函数，则由这种罚函数构成的稀疏模型能够实现双层的变量选择. 由这种层次罚函数构成的具有双层稀疏性的结构稀疏模型有很多：$L_1$ 范数组 SCAD 模型[24]的罚函数内部采用了 $L_1$ 范数罚，外部采用了 SCAD 罚；$L_1$ 范数组 MC 模型[24]的罚函数内部采用了 $L_1$ 范数罚，外部采用了 MC 罚；$L_1$ 范数组桥模型[24]的罚函数内部采用了 $L_1$ 范数罚，外部采用了桥罚；指数组套索（Group Exponential Lasso）[37]内部采用了 $L_1$ 范数罚，外部采用了 Exponential 罚；复合组 MC 模型[24]的罚函数内部和外部同时采用了 MC 罚，其相对于前述几种内部采用 $L_1$ 范数罚的双层稀疏模型具有组向量选择和组内分量选择一致性；使用内外桥罚的复合组桥罚也具有组向量选择和组内分量选择一致性[39].

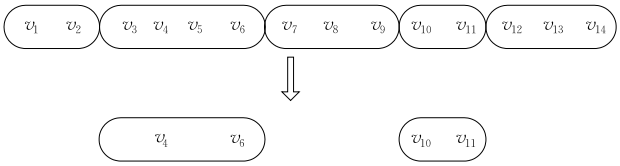



图 7 双层结构稀疏模型的稀疏化效果

### 3.2 多层稀疏模型

除双层稀疏模型外，Gao 和 Chia 等人[40]提出一种多层稀疏模型. 在图像分类和图像标注中，他们提出一种多层组稀疏编码方法，该方法的稀疏结构分为三层：实例层(instance layer)、基于类的分组层(class-based group layer)和基于标注的子组层(tag-based subgroup layer). 假设图像共有 $M$ 个类标签，第 $m$ 类的全部图像样本组成矩阵 $\boldsymbol{X}_m$，其中 $m\in\{1,2,\cdots,M\}$. 第 $m$ 类的全部图像又进一步被划分为 $G_m$ 个基于标注的子组 $X_{m1}, X_{m2},\cdots, X_{mg},\cdots, X_{mG_m}$，其中 $g\in\{1,2,\cdots,G_m\}$. 令 $X_{mg}^k\in\boldsymbol{R}^{d\times 1}$ 表示基于标注的子组中的第 $k$ 个元素，$k\in\{1,2,\cdots,M_{mg}\}$，其中 $M_{mg}$ 为第 $g$ 个基于标注的子组 $X_{mg}$ 中的元素数目，$d$ 表示每个图像的特征维数. 总之，令 $\boldsymbol{X}$ 表示全部图像对应的数据矩阵，且 $\boldsymbol{X}_{mg}\in\boldsymbol{R}^{d\times M_{mg}}$，$\boldsymbol{X}_i\in\boldsymbol{R}^{d\times\sum_g N_{mg}}$，$\boldsymbol{X}\in\boldsymbol{R}^{d\times\sum_m\sum_g M_{mg}}$，则 Gao 和 Chia 等人提出的多层组稀疏编码方法为求解如下的优化问题：

$$\min_{\boldsymbol{V}}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{V}\|_F^2+\lambda\|\boldsymbol{V}\|_1+\sum_{m=1}^{M}w_m\|\boldsymbol{V}_m\|_p+\sum_{m=1}^{M}w_m\sum_{g=1}^{G_m}\gamma_{mg}\|\boldsymbol{V}_{mg}\|_p,$$

其中 $\|\cdot\|_F$ 表示 Frobenius 范数，$\|\boldsymbol{V}\|_1$ 表示矩阵 $\boldsymbol{V}$ 的全部元素的绝对值之和，而 $\boldsymbol{V}_{mg}\in\boldsymbol{R}^{M_{mg}\times 1}$，$\boldsymbol{V}_m\in\boldsymbol{R}^{\sum_g M_{mg}\times 1}$ 和 $\boldsymbol{V}\in\boldsymbol{R}^{\sum_m\sum_g M_{mg}\times 1}$ 为对应的模型向量的系数矩阵. Gao 和 Chia 等人提出的多层稀疏模型充分挖掘了图像类标签(class label)和标注(tag)之间的依赖关系，相应地引入了基于标注的子组层这一正则化项，构造出了具有多层组稀疏结构的编码方法，可同时实现图像标注和图像分类.

## 4 树结构稀疏模型

### 4.1 树组套索模型

树组套索假定特征组结构之间满足偏序关系. 例如，图像处理中各像素间的关系往往为树结构或森林结构[41]；再比如，在基因表达分析中各个基因之间往往具有偏序关系，即某基因能否表达取决于另一基因是否已经表达. 层次稀疏组结构假定变量组成树结构或森林结构，树结构或森林结构的节点上变量设置为 0 时，所有其子节点变量也都置为 0，或者说，变量属于最终的模型向量的支集只当其所在树或森林中的祖先节点属于最终的模型向量的支集. 当处理这种数据时，需要充分利用树结构作为先验信息[42]. 假定 $g_1,\cdots,g_J$，$g_j\subseteq\{1,\cdots,P\}$，$G=\{g_j\,|\,j=1,\cdots,J\}$，$\boldsymbol{\beta}_{g_j}$ 表示组 $g_j$ 的子模型向量，$g_j\in G$ 和 $g_{j^*}\in G$ 为任意的两个组，满足关系：

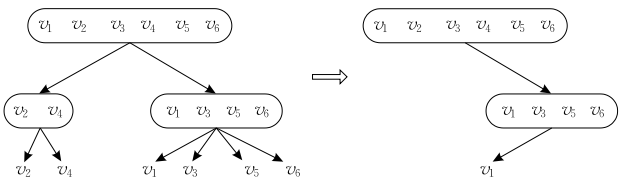
$$(g_j\cap g_{j^*}\neq\phi)\Rightarrow(g_j\subseteq g_{j^*}\cup g_{j^*}\subseteq g_j) \quad (20)$$

$$\bigcup_{g_j\in G}g_j=\{1,\cdots,P\} \quad (21)$$

树组套索目标函数为[43]

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2 \quad (22)$$

这里 $\lambda\geq 0$ 为正则化因子，$w_{g_j}$ 为组 $g_j$ 的权，$G$ 具有树结构. 实际上，树结构是重叠组结构的一个特例，因而树组套索也是重叠组套索的一个特例，树结构在重叠组结构的基础上附加如下三个条件：同一层中的各组节点之间不包含共有的变量；树中子节点的索引集是父节点索引集的子集；树中父节点的索引集为树中其子节点索引集的覆盖集. 图 8(a)为 6 个节点的树结构模型，图 8(b)为树组套索的选择效果，其中椭圆形内部的变量属于一个组. 假设树组套索的选择结果为选中了组$\{v_1\}$，而组$\{v_2,v_4\}$、$\{v_3\}$、$\{v_5\}$和$\{v_6\}$被丢弃了，如图 8 所示. 由于选中了$\{v_1\}$，故其全部父组$\{v_1,v_2,v_3,v_4,v_5,v_6\}$和$\{v_1,v_3,v_5,v_6\}$也就被选中了；同时，由于$\{v_2,v_4\}$被丢弃了，那么其所有的子组$\{v_2\}$和$\{v_4\}$也就被丢弃了. 树组套索的选择效果的特点体现在如下两个方面：若某个组(节点)被树组套索选中，那么该组(节点)的全部父组(父节点)也被选中；若某个组(节点)被树组套索丢弃，那么该组(节点)的全部子组(子节点)也被丢弃.




(a) 6个节点的树结构 (b) 树组套索的选择效果

图 8 树组套索的稀疏化效果

### 4.2 多输出树组套索模型

在多输出变量情形下，假设输出变量具有分层的树结构，其中叶子节点对应单个输出变量，而内部节点对应一组输出变量，子节点只与包含子节点元素的父节点有边连接，根节点为所有叶子节点元素组

成的集合.将此树结构作为先验信息,可以得到输出具有树结构的多输出树组套索.多变量线性模型为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{B}+\boldsymbol{W} \tag{23}$$

这里,$\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$为输入矩阵,$\boldsymbol{Y}\in\boldsymbol{R}^{N\times K}$为输出变量矩阵,$\boldsymbol{B}\in\boldsymbol{R}^{P\times K}$为模型矩阵,$\boldsymbol{W}\in\boldsymbol{R}^{N\times K}$表示噪声矩阵.

假设多个输出变量之间的关系为树结构 $T$.令 $G=\{g_j\mid j=1,\cdots,J\}$, $g_1,\cdots,g_J$中的每一个下标子集 $g_j\subseteq\{1,\cdots,P\}$满足 $\bigcup\limits_{g_j\in G}g_j=\{1,\cdots,P\}$, 如图 9 和图 10 所示.图中,叶子节点对应单个输出变量,内部节点对应下标子集 $g_j\subseteq\{1,\cdots,P\}$所确定的输出变量集,节点 $g_j\in G$ 的权为 $w_{g_j}$.输出变量具有树结构的多输出树组套索问题为[44]

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2 \tag{24}$$

这里,$\lambda\geq0$,$\|\cdot\|_F$表示Frobenius范数,$\boldsymbol{\beta}_{g_j}$表示组(节点)$g_j$对应的子模型向量.正则化项 $\sum\limits_{p\in\{1,\cdots,P\}}\sum\limits_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2$ 包含$L_{2,1}$运算$\sum\limits_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2$和外部的 $L_1$运算 $\sum\limits_{p\in\{1,\cdots,P\}}$.内部的 $L_{2,1}$范数运算$\sum\limits_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2$表示对系数矩阵 $\boldsymbol{B}$ 的第 $p$ 行进行如下操作:(1)将第 $p$ 行的全部系数进行分组,所划分的组具有如图 8 所示的树结构.其中罚函数中的$\boldsymbol{\beta}_{g_i}^p$表示系数矩阵 $\boldsymbol{B}$ 的第 $p$ 行中第 $g_i$个组的系数;(2)对第 $p$ 行中各组分别单独进行 $L_2$范数运算;(3)在上述组水平上执行 $L_1$范数运算,于是导致系数矩阵 $\boldsymbol{B}$第 $p$ 行内某些组为零,造成行内的组稀疏.由于系数矩阵 $\boldsymbol{B}$ 的第 $p$ 行全部表示输出关于第 $p$ 个变量的系数向量,代表了输出之间的关系,因此第(1)步中第 $p$ 行内组之间的树结构关系即对应 $K$ 个输出之间的树结构关系,并且第 $p$ 行内的组稀疏就相当于从 $K$ 个输出角度而言的稀疏.外部的 $L_1$范数运算 $\sum\limits_{p\in\{1,\cdots,P\}}$ 表示将系数矩阵 $\boldsymbol{B}$ 的一行作为一个组,在此组(行)水平上进行 $L_1$范数运算,导致某些行的系数全部为零.由于系数矩阵 $\boldsymbol{B}$ 的列表示同一输出关于全部变量的系数,因此从 $P$ 个变量角度而言实现了稀疏.因此,树组套索能够同时实现从变量角度而言的稀疏和从输出角度而言的稀疏.以较简单的 3 个输出变量的情况来进行说明.此时系数矩阵为

$$\boldsymbol{B}=\begin{pmatrix}\beta_{11}&\beta_{12}&\beta_{13}\\\beta_{21}&\beta_{22}&\beta_{23}\\\beta_{31}&\beta_{32}&\beta_{33}\end{pmatrix} \tag{25}$$

其第 1 行对应的分组情况如图 9 所示,其中一个椭圆对于一个组,第一行对应的输出变量的树结构关系如图 10 所示.图 9 中的组即为图 10 中的节点,两者均用 $g_j$表示.

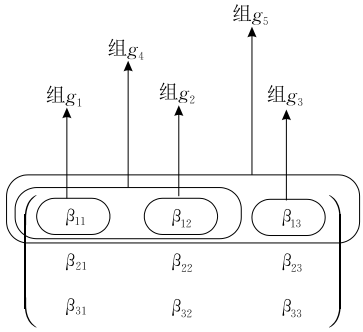



图 9　系数矩阵第 1 行的分组情况

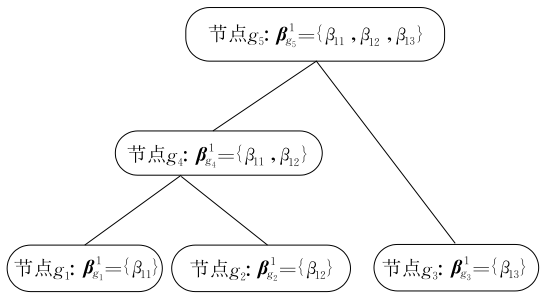



图 10　输出变量的树结构

### 4.3　输入与输出联立 $L_1/L_\infty$ 罚套索

Turlach 等人[45]提出使用输入预测变量的同一子集同时预测几个目标响应变量的模型,表示为联立变量选择(Simultaneous Variable Selection,SVS).SVS 把套索用于式(23)表示的多输入输出模型,在式(23)中

$$\boldsymbol{X}=\begin{pmatrix}X_{11}&\cdots&X_{1P}\\\vdots&\ddots&\vdots\\X_{N1}&\cdots&X_{NP}\end{pmatrix}\in\boldsymbol{R}^{N\times P},$$

$$\boldsymbol{Y}=\begin{pmatrix}Y_{11}&\cdots&Y_{1K}\\\vdots&\ddots&\vdots\\Y_{N1}&\cdots&Y_{NK}\end{pmatrix}=(Y^1,\cdots,Y^K)=\{Y_{il}\}\in\boldsymbol{R}^{N\times K},$$

$$\boldsymbol{B}=\{\beta_{jl}\}=(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_P^{\mathrm{T}})^{\mathrm{T}}=(\beta^1,\cdots,\beta^K)\in\boldsymbol{R}^{P\times K},$$

$Y^l\in\boldsymbol{R}^N$,$\boldsymbol{\beta}_j\in\boldsymbol{R}^K$,$\beta^l\in\boldsymbol{R}^P$.SVS 结构稀疏模型假设 $N$ 个预测变量中每一个预测变量的特征分量 $X_{i1},\cdots,X_{iP}$中只有很少的子集变量与输出变量 $Y_{i1},\cdots,Y_{iK}$相关,输入与输出联立 $L_1/L_\infty$罚套索(Simultaneous $L_1/L_\infty$ Penalty LASSO)问题为

$$\min_{\beta_{11},\cdots,\beta_{PK}}\sum_{l=1}^{K}\sum_{i=1}^{N}(Y_{il}-\sum_{j=1}^{P}X_{ij}\beta_{jl})^2$$

$$\text{s.t.}\ \sum_{j=1}^{P}\max(|\beta_{j1}|,\cdots,|\beta_{jK}|)\leq t$$

其等价形式为

$$\min_{\boldsymbol{B}}\sum_{l=1}^{K}\|Y^l-\boldsymbol{X}\boldsymbol{\beta}^l\|_2^2+\lambda\sum_{j=1}^{P}\|\boldsymbol{\beta}_j\|_\infty,$$

这里 $\|\boldsymbol{\beta}_j\|_\infty=\max(|\beta_{j1}|,\cdots,|\beta_{jK}|)$ 为 $L_1/L_\infty$ 范数，这里 $L_\infty$ 罚只依赖于 $\boldsymbol{B}$ 每一行最大的系数，对别的非 0 系数不执行罚，组内没有稀疏化，如果某个预测变量被选中($\boldsymbol{B}$ 的某行)，则执行选中 $K$ 个非 0 系数的回归模型，因此 $L_1/L_\infty$ 范数与 $L_1/L_\infty$ 块范数一样，实现组之间稀疏，而组内不稀疏的模型.

## 5 图结构稀疏模型

稀疏正则化思想还被应用于概率图模型的学习中. 概率图模型学习的本质任务为对其 precision 矩阵的学习，于是利用 $L_1$ 范数罚对概率图模型进行稀疏学习，这种方法即所谓的图套索(Graphical Lasso)[44,46]：

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}>0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda\|\boldsymbol{\Theta}\|_1 \tag{26}$$

其中 $\boldsymbol{S}$ 为经验协方差矩阵，$\boldsymbol{\Theta}=\Sigma^{-1}$ 为概率图模型的 precision 矩阵，precision 矩阵中的元素是否为零对应着概率图模型中两个节点 $i$ 和 $j$ 之间是否存在一条边连接，$tr(\cdot)$ 表示矩阵的迹，$\|\boldsymbol{\Theta}\|_1=\sum_{j=1}^{P}\sum_{i=1}^{P}|\Theta_{ij}|$. 利用 $L_1$ 范数罚对概率图模型进行稀疏学习为最简单的情形，后来又有学者陆续提出了新的正则化项，例如 $L_{\infty,1}$ 范数组结构图套索[44]：

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}>0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda\sum_{g=1}^{m}\max\{|\Theta_{ij}|:(i,j)\in G_g\} \tag{27}$$

其中 $\sum_{g=1}^{m}\max\{|\Theta_{ij}|:(i,j)\in G_g\}$ 为 $L_{\infty,1}$ 范数罚，表示先对分组 $G_g$ 中的元素进行 $L_\infty$ 范数运算，然后再横跨全部组进行 $L_1$ 范数运算. 再例如 $L_{2,1}$ 范数组结构图套索[46]：

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}>0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\sum_{g=1}^{m}\|\{|\Theta_{ij}|:(i,j)\in G_g\}\|_2 \tag{28}$$

其中 $\sum_{g=1}^{m}\|\{|\Theta_{ij}|:(i,j)\in G_g\}\|_2$ 为 $L_{2,1}$ 范数罚，表示先对分组 $G_g$ 中的元素进行 $L_2$ 范数运算，然后再横跨全部组进行 $L_1$ 范数运算. 这两种图套索具有令 precision 矩阵中元素成组地进行稀疏的效果.

## 6 结构化稀疏字典学习

结构化稀疏字典学习分类如图 11 所示. 信号处理和统计学中稀疏编码或者字典学习的一个起源来自于矩阵因式化，即用两个未观测的矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 表示一个已知矩阵 $\boldsymbol{X}\approx\boldsymbol{A}\boldsymbol{D}$，PCA 是这一思想的典型应用[47]，这里 $\boldsymbol{D}$ 的列向量为字典元素，$\boldsymbol{A}$ 为相应的系数矩阵. 但是使用预先固定的字典不一定是特定信号最优的字典，不能产生适当稀疏的信号表示. 稀疏编码假定数据由一些字典元素线性组合形成，最近，这些思想在神经科学，生物信息学，计算机视觉中的图像解噪、超分辨率分析、图像重构以及用可操纵小波去马赛克中被广泛应用. 字典学习的另一个起源来自于建立自然图像统计结构与哺乳动物的视觉皮层区域 $V1$ 神经元之间关系[48]. 由于用整个矩阵表示自然图像时矩阵过大无法计算，而且实际自然图像样本个数 $n$ 非常大，而信号的维数 $m$ 相比较样本个数 $n$ 而言非常小，而且常常选择字典元素的个数 $p$ 大于 $m$，此时称为过完备(overcomplete)，所以，神经科学研究中，Olshausen 和 Field 提出学习训练数据上的的图像块的自适应字典稀疏码[47,48]，典型的学习场景中，如样本个数为 $n=100\,000$、字典元素的个数为 $p=256$ 以及图像块选为 $m=16\times16$，Olshausen 研究显示稀疏字典能够自动发现自然图像统计结构与哺乳动物的视觉皮层区域 $V1$ 神经元之间关系[47,48]，在图像重构邻域，稀疏字典显著优于使用固定的字典元素如曲线小波(curvelets)、轮廓小波(contourlets)或带通小波 (bandlets)的重构效果. 不用任何先验假设，只用稀疏模型假设就能够产生类似于空间局部有向基函数 Gabor 小波的字典元素[49,50].

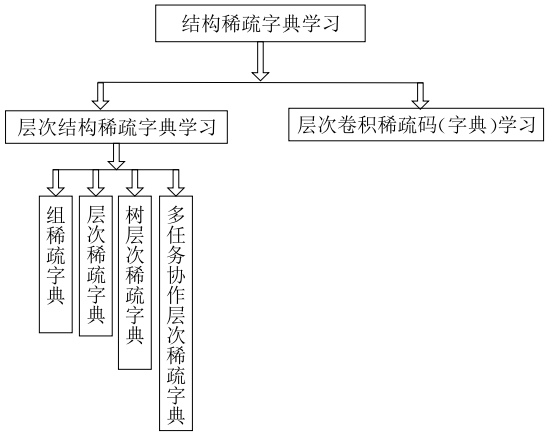



图 11 结构化稀疏字典学习

给定 $n$ 个训练样例 $\boldsymbol{X}=(x_1,\cdots,x_n)\in\boldsymbol{R}^{m\times n}$，$x_i\approx\boldsymbol{D}\boldsymbol{a}_i=\sum_{j=1}^{p}a_{ij}d_j\in R^m$，$\boldsymbol{D}=(d_1,\cdots,d_p)\in\boldsymbol{R}^{m\times p}$，$\boldsymbol{A}=(a_1,\cdots,a_n)\in\boldsymbol{R}^{p\times n}$，稀疏字典学习可归结为以下优化问题[51-53]

$$\min_{\boldsymbol{D}\in\Xi,\boldsymbol{A}\in\boldsymbol{R}^{p\times N}}\sum_{i=1}^{n}\frac{1}{2}\|x_i-\boldsymbol{D}\boldsymbol{a}_i\|_2^2+\lambda\psi(\boldsymbol{a}_i),$$
$$\Xi\equiv\{\boldsymbol{D}\in\boldsymbol{R}^{m\times p}:\forall j\ \|d_j\|_2\leq 1\}\tag{29}$$

这里，$\psi(\boldsymbol{a}_i)$为诱导稀疏的正则化项. 上述问题也可以等价写为

$$\min_{\boldsymbol{D}\in\Xi,\boldsymbol{A}\in\boldsymbol{R}^{p\times n}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2+\lambda\psi(\boldsymbol{A}),\psi(\boldsymbol{A})=\sum_{i=1}^{n}\psi(\boldsymbol{a}_i),$$

这体现了字典的因式分解观点形式. 这种表示得到字典学习与其他无监督学习如非负矩阵因式化和聚类等之间的关联关系.

除了稀疏假设外，实际的数据的支集本身具有结构约束，例如在网络异常检测时，可以假设异常行为从某些网络节点产生，形成星形模式子网结构. 在基因表达研究中，某些不同个体的基因组对同一类型疾病具有某一相同的生物学机制，在这一情况，基因表达矩阵的支集应该是其子矩阵，能否用这些结构信息进一步地增加稀疏模型对真实支集估计算法的性能是值得研究的.

### 6.1　层次结构稀疏字典学习

(1) 组稀疏字典

组稀疏字典罚为

$$\psi(\boldsymbol{a}_i)=\sum_{j=1}^{J}\|\boldsymbol{a}_i[g_j]\|_2,$$

这里，$\bigcup_{g_j\in G}g_j=\{1,\cdots,p\}$的子集 $g_j\subseteq\{1,\cdots,p\}$组成组结构 $G=\{g_j\,|\,j=1,\cdots,J\}$，组稀疏字典罚选择 $\boldsymbol{a}_i=(a_{i1},\cdots,a_{ip})\in R^p$的某些个分量组成组子向量[54].

(2) 层次稀疏字典

层次稀疏字典罚为

$$\psi(\boldsymbol{a}_i)=\lambda_2\sum_{j=1}^{J}\|\boldsymbol{a}_i[g_j]\|_2+\lambda_1\|\boldsymbol{a}_i\|_1$$

层次稀疏字典罚不但用$\sum_{j=1}^{J}\|\boldsymbol{a}_i[g_j]\|_2$罚项实现组稀疏字典，而且还用$\|\boldsymbol{a}_i\|_1$罚项实现组内稀疏功能[55].

(3) 树层次稀疏字典

假定 $p$ 个字典元素中每个字典元素对应树中的某个节点，$p$ 个字典元素存在层次组稀疏树关系，树结构由先验知识人为确定，此时，罚函数 $\psi(\boldsymbol{a}_i)$选为层次组套索罚

$$\psi(\boldsymbol{A})=\sum_{i=1}^{n}\sum_{g_j\in G}^{J}\|\boldsymbol{a}_i[g_j]\|_q,$$

这里，$\|\cdot\|_q$范数可选为 $L_\infty$ 或 $L_2$ 范数. $\bigcup_{g_j\in G}g_j=\{1,\cdots,p\}$的子集 $g_j\subseteq\{1,\cdots,p\}$组成组结构 $G=\{g_j\,|\,j=1,\cdots,J\}$，每个组包含树中的某个节点及其在树中的子节点，$\boldsymbol{a}_i[g_j]$表示 $\boldsymbol{a}_i$对应 $g_j$组索引集的子模型向量[56].

实验发现，字典元素自然地把图像块组合成树结构，图像低频区域位于树的根节点，图像高频区域位于叶子节点. 树中的父节点与其子节点具有更强的相关关系，子节点代表的图像与父节点图像更相似，但是子节点代表的图像比父节点代表的图像具有更高的频率并且表示的图像稍微有些变化[57].

(4) 多任务协作层次稀疏字典

很多情况下，我们希望 $n$ 个训练样例$\{x_1,\cdots,x_n\}$共享字典同样的子向量，即对于每一个 $x_i\in\{x_1,\cdots,x_n\}$，$\boldsymbol{a}_i$的非 0 分量集合是一样的. 该问题为多任务协作稀疏字典问题，此时 $\psi(\boldsymbol{A})=\sum_{k=1}^{p}\|\boldsymbol{a}^k\|_2$，这里 $\boldsymbol{a}^k\in\boldsymbol{R}^n$为矩阵 $\boldsymbol{A}$ 的第 $k$ 行，$\boldsymbol{a}^k$的分量相应于所有 $x_i\in\{x_1,\cdots,x_n\}$的第 $k$ 个原子(atom). 使用组套索得到多任务协作组稀疏字典问题，相应的罚函数为 $\psi(\boldsymbol{A})=\sum_{j=1}^{J}\|\boldsymbol{A}^{[g_j]}\|_F$，这里 $\boldsymbol{A}^{[g_j]}$为所有属于 $g_j$的行形成的子矩阵.

多任务协作层次稀疏字典罚函数为

$$\psi(\boldsymbol{a}_i)=\lambda_2\sum_{j=1}^{J}\|\boldsymbol{A}^{[g_j]}\|_F+\lambda_1\sum_{i=1}^{n}\|\boldsymbol{a}_i\|_1.$$

当设置 $\lambda_1=0$，得到多任务协作组稀疏字典问题，当设置 $\lambda_2=0$，可获得每个 $x_i\in\{x_1,\cdots,x_n\}$的套索解. $\{x_1,\cdots,x_n\}$的多任务协作组稀疏字典表示形式不依赖于组内分量的特定的表示和大小，组内分量合起来，才决定$\{x_1,\cdots,x_n\}$的字典，多任务协作层次稀疏字典罚函数能发现对应于同一组结构上述的$\{x_1,\cdots,x_n\}$表示形式.

### 6.2　层次卷积稀疏码(字典)学习

与前述图像块学习相反，Zhu 等人[58]提出学习整个图像的卷积稀疏字典方法，整个图像的卷积稀疏字典由在图像的各个位置上的小的字典元素线性组合得到.

Zeiler 等人[59,60]提出了层次卷积稀疏码模型(hierarchical convolutional sparse coding model). 假定输入图像 $X^i$有 $K_0$个颜色通道 $X_1^i,\cdots,X_{K_0}^i$，假定第 $c$ 个通道 $\boldsymbol{X}_c^i$为 $K_1$个隐特征 $z_k^i$与过滤子 $\boldsymbol{f}_{k,c}$卷积形成

$$\boldsymbol{X}_c^i=\sum_{k=1}^{K_1} z_k^i \otimes \boldsymbol{f}_{k,c},$$

这里，$\boldsymbol{X}_c^i \in \boldsymbol{R}^{N_r \times N_c}$，$\boldsymbol{f}_{k,c} \in \boldsymbol{R}^{H \times H}$，共有 $K_1=(N_r+H-1)\times(N_c+H-1)$ 个隐特征 $z_k^i$. 卷积稀疏码模型目标函数为

$$C_1(X^i)=\frac{\lambda}{2}\sum_{c=1}^{K_0}\left\|\sum_{k=1}^{K_1} z_k^i \otimes \boldsymbol{f}_{k,c}-\boldsymbol{X}_c^i\right\|_2^2+\sum_{k=1}^{K_1}\|z_k^i\|_1.$$

卷积稀疏码模型学习时，对图像集 $\boldsymbol{X}=\{X^1,\cdots,X^I\}$，每一个隐特征映射 $f=(f_{1,1},\cdots,f_{K_1,K_0})$ 和过滤子 $z=(z_1^i,\cdots,z_{(N_r+H-1)\times(N_c+H-1)}^i)$，求解 $\arg\min_{f,z} C_1(X)$ 学习最优的隐特征映射 $f$ 和过滤子 $\boldsymbol{z}$，$\boldsymbol{X}$ 里的每个图像具有自己的隐特征映射函数，但是 $\boldsymbol{X}$ 中的所有图像公用同样的过滤子. $f$ 把上述特征提取过程分层，把第 $l$ 层的特征映射 $z_{k,l}^i$ 作为第 $l+1$ 层的输入，$l$ 层的图像输入为具有 $K_{l-1}$ 个颜色通道的 $l-1$ 层的特征映射 $z_{k,l-1}^i$，第 $l$ 层的目标函数为

$$C_l(X^i)=\frac{\lambda}{2}\sum_{i=1}^{I}\sum_{c=1}^{K_{l-1}}\left\|\sum_{k=1}^{K_l}\boldsymbol{g}_{k,c}^l(z_{k,l}^i \otimes f_{k,c}^l)-z_{c,l-1}^i\right\|_2^2+\sum_{i=1}^{I}\sum_{k=1}^{K_l}\|z_{k,l}^i\|_1,$$

$z_{c,l-1}^i$ 为前一层的特征映射，二值矩阵 $\boldsymbol{g}_{k,c}^l$ 决定两层之间的特征映射之间的连接关系，二值矩阵 $\boldsymbol{g}_{k,c}^l$ 决定 $z_{k,l}^i$ 是否连接到 $z_{c,l-1}^i$，第一层的二值矩阵 $\boldsymbol{g}_{k,c}^1$ 为全 1 矩阵.

相似于文献[58]提出的卷积稀疏码方法，Bo 等人在文献[61]提出组合具有不同层数的多个网络，使用层次匹配追踪(hierarchical matching pursuit)方法构造多路径稀疏码(Multipath sparse coding)，对每个图像构造多层特征映射序列，但是编码过程对卷积稀疏码进行了简化，简化的部分主要有：所有图像块的特征映射编码相互独立，使用贪心正交匹配过程构造编码，Bo 等人的方法在图像识别任务中大获成功. 另外，只具有单层的简化版的此类层次卷积稀疏码也显示能够有效地替代梯度直方图构造的低水平特征[62]，可用于计算机视觉对象检测任务[63,64].

另外，类似于独立成分分析形式，Kavukcuoglu 等人[65]提出具有两维网格结构的字典，与层次结构的字典学习类似，字典元素组成网格状，字典元素之间可定义近邻关系，例如可以把 $p$ 个字典元素组成 $\sqrt{p}\times\sqrt{p}$ 循环网格，具有 $3\times3$，$4\times4$ 的空间字典元素近邻元素重叠组成 $p$ 个组.

Garrigues 和 Olshausen 在文献[66]中提出可以在近邻变量组上定义隐变量概率模型构造混合稀疏码，Gregor 等人在文献[67]提出字典元素之间存在禁止近邻关系的网状结构字典学习模型.

## 7 结构稀疏模型的统一框架

### 7.1 结构稀疏化与贝叶斯理论之间的联系

贝叶斯理论认为，可以把组套索表示为贝叶斯最大后验估计问题[68,69]，其中组套索的损失函数项对应贝叶斯最大后验估计中的似然函数，组套索的罚函数对应贝叶斯最大后验估计中的先验概率. 实际上，由结构风险最小化理论可知，其他的结构稀疏模型也可被视为将变量中存在的结构作为先验信息的贝叶斯最大后验估计，即结构稀疏模型具有贝叶斯形式的统一框架. 贝叶斯理论揭示了结构稀疏模型的拓展方向，改变似然函数，将其向其他统计模型拓展，或者改变其先验概率，设计新的正则化项.

### 7.2 结构稀疏化与编码复杂度之间的联系

Huang 等人[70]利用编码复杂度理论观点来审视稀疏模型和结构稀疏模型，并将其称作编码复杂度正则化(Coding Complexity Regularization)方法. Huang 等人将非结构化的稀疏性、组结构稀疏性、树结构稀疏性和图结构稀疏性等纳入了编码复杂度正则化的统一框架. 另外，Huang 等人所做的工作的一个重要意义在于其利用所提出的编码复杂度正则化理论揭示了在何种情况下结构稀疏模型优于非结构化的稀疏模型.

### 7.3 结构稀疏化与子模函数之间的联系

Bach 等人[71]利用子模函数建立了结构稀疏模型的统一框架. $L_0$ 范数表示模型向量中非零元素的个数，其等价于模型向量支撑集的势(cardinality)，其是一种集合函数并且是实现稀疏性的最直观方法，但由于对其难以求解从而利用其凸包络 $L_1$ 范数代替. Bach 等人考虑了更广泛的集合函数，假设集合函数 $F(\cdot)$ 为任意一非降的子模函数，$f(\cdot)$ 为其 Lovasz 展开(Lovasz extension)，$|\boldsymbol{\beta}|$ 表示模型向量的每个元素均取绝对值后构成的新向量，$\mathrm{supp}(\boldsymbol{\beta})$ 表示模型向量 $\boldsymbol{\beta}$ 的支撑集，同时指出并证明了 $f(|\boldsymbol{\beta}|)$ 是一种范数并且是 $F(\mathrm{supp}(\boldsymbol{\beta}))$ 的凸包络. 显然，要想得到 $F(\mathrm{supp}(\boldsymbol{\beta}))$ 的凸包络，对其进行 Lovasz 展开即可. Bach 等人指出非重叠组结构稀疏性和重叠组结构稀疏性等很多结构化稀疏性都可纳入这个框架.

### 7.4 通过引入辅助向量构造出的结构稀疏模型统一框架

Micchelli 等人[72]引入一辅助向量 $\boldsymbol{\gamma}\in\boldsymbol{R}^P$ 构造了结构稀疏模型的一种统一框架：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\Omega(\boldsymbol{\beta}|\boldsymbol{\Lambda}) \tag{30}$$

其中

$$\Omega(\boldsymbol{\beta}|\boldsymbol{\Lambda})=\inf\{\Gamma(\boldsymbol{\beta},\gamma):\gamma\in\boldsymbol{\Lambda}\} \tag{31}$$

$$\Gamma(\boldsymbol{\beta},\gamma)=\frac{1}{2}\sum_{p=1}^{P}\left(\frac{\beta_p^2}{\gamma_p}+\gamma_p\right) \tag{32}$$

向量集 $\boldsymbol{\Lambda}$ 的选定非常重要，选择不同的向量集 $\boldsymbol{\Lambda}$ 就会构造出不同的结构稀疏模型. 例如，Micchelli 等人构造了一种结构稀疏模型，这种结构稀疏模型的罚函数叫做箱形罚，由其得到的模型向量 $\boldsymbol{\beta}$ 的各分量均处于规定的区间中，他们指出此时选定 $\boldsymbol{\Lambda}=\{(\gamma_p:p\in\{1,\cdots,P\}:\gamma_p\in[a_p,b_p])\}$ 则可实现这种结构稀疏化效果；Micchelli 等人还构造了一种叫做楔形罚的罚函数，由楔形罚构成的结构稀疏模型得到的模型向量各分量的大小是非递增的，他们指出选定 $\boldsymbol{\Lambda}=\{(\gamma_p:p\in\{1,\cdots,P\}:\gamma_p\geq\gamma_{p+1})\}$ 即可实现这种结构稀疏化效果. Micchelli 等人提出的这种框架有力地对结构稀疏模型进行了概括，可通过选定不同的向量集 $\boldsymbol{\Lambda}$ 来构造出具有不同特性的结构稀疏模型.

## 8 结构稀疏模型的统计特性

参数估计一致性、变量选择一致性和 oracle 性质是结构稀疏模型中被讨论最多的统计特性，而这些统计特性当前主要出现在组结构稀疏模型中，树结构和图结构等结构稀疏模型的统计特性当前尚少有研究，因此本节以组结构稀疏模型为例探讨实现这些统计特性的附加条件. 参数估计一致性是指对于任意正数 $\varepsilon^*$ 有 $\lim_{N\to\infty}P(\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2<\varepsilon^*)=1$ 成立，其中 $N$ 是样本个数. 变量选择一致性是指 $\lim_{N\to\infty}P(\{j:\hat{\boldsymbol{\beta}}_j\neq0\}=\{j:\boldsymbol{\beta}_j\neq0\})=1$，而 oracle 性质[16]包括两方面：

(1) 变量选择一致性：

$$\lim_{N\to\infty}P(\{j:\hat{\boldsymbol{\beta}}_j\neq0\}=\{j:\boldsymbol{\beta}_j\neq0\})=1;$$

(2) 参数估计渐近正态性：令 $\boldsymbol{A}^*=\{j:\hat{\boldsymbol{\beta}}_j\neq0\}$，则当 $N\to\infty$ 时有

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\boldsymbol{A}^*}-\boldsymbol{\beta}_{\boldsymbol{A}^*})\to\boldsymbol{N}(\boldsymbol{0},\Sigma).$$

与 Tibshirani 的套索类似，组套索的估计值一般不具有变量选择一致性，或者说只有在很强的假设条件下才具有变量选择一致性. 这是因为组套索对于各个组对应的子模型向量 $\boldsymbol{\beta}_j$ 的惩罚程度不随子模型向量 $\boldsymbol{\beta}_j$ 的模的大小而自适应地改变，因此会过度缩小模较大的子模型向量的值，导致对模较大的子模型向量产生有偏估计. 为了补偿对模较大的子模型向量的过度缩小，组套索会倾向于将一些不必要的组也选择出来，即组套索具有对组过选择的特性. 套索模型的参数估计一致性、变量选择一致性和 oracle 性质的讨论需要附加假设条件，这些假设条件主要有不可表示条件[73](irrepresentable condition)、稀疏 Riesz 条件[74](sparse Riesz condition)和限制特征值条件[75](restricted eigenvalue condition)等，类似地，在讨论组套索等结构稀疏模型的参数估计一致性、变量选择一致性和 oracle 性质时也要附加上述 3 个假设条件.

不可表示条件是组套索满足变量选择一致性的必要条件，分为强不可表示条件和弱不可表示条件. Bach 等人[73]将套索模型中的不可表示条件推广到组套索中，并指出组套索在满足不可表示条件时具有变量选择一致性. 已知 $\widetilde{B}=\{j:\boldsymbol{\beta}_j\neq0\}$，$\widetilde{B}^C$ 表示 $\widetilde{B}$ 的补集，则关于组套索的设计矩阵的强不可表示条件和弱不可表示条件分别为

$$\max_{j^*\in\widetilde{B}^C}\frac{1}{\sqrt{d_{j^*}}}\left\|\Sigma_{\boldsymbol{X}_{j^*}\boldsymbol{X}_{\widetilde{B}}}\Sigma_{\boldsymbol{X}_{\widetilde{B}}\boldsymbol{X}_{\widetilde{B}}}^{-1}\operatorname{diag}\left(\frac{\sqrt{d_j}}{\|\boldsymbol{\beta}_j\|_2}\right)\boldsymbol{\beta}_{\widetilde{B}}\right\|<1 \tag{33}$$

$$\max_{j^*\in\widetilde{B}^C}\frac{1}{\sqrt{d_{j^*}}}\left\|\Sigma_{\boldsymbol{X}_{j^*}\boldsymbol{X}_{\widetilde{B}}}\Sigma_{\boldsymbol{X}_{\widetilde{B}}\boldsymbol{X}_{\widetilde{B}}}^{-1}\operatorname{diag}\left(\frac{\sqrt{d_j}}{\|\boldsymbol{\beta}_j\|_2}\right)\boldsymbol{\beta}_{\widetilde{B}}\right\|\leq1 \tag{34}$$

其中

$$\Sigma_{\boldsymbol{X}_{j^*}\boldsymbol{X}_{\widetilde{B}}}=E(\boldsymbol{X}_{j^*}\boldsymbol{X}_{\widetilde{B}})-E(\boldsymbol{X}_{j^*})E(\boldsymbol{X}_{\widetilde{B}}),$$

$$\Sigma_{\boldsymbol{X}_{\widetilde{B}}\boldsymbol{X}_{\widetilde{B}}}=E(\boldsymbol{X}_{\widetilde{B}}\boldsymbol{X}_{\widetilde{B}})-E(\boldsymbol{X}_{\widetilde{B}})E(\boldsymbol{X}_{\widetilde{B}}),$$

$\operatorname{diag}\left(\frac{\sqrt{d_j}}{\|\boldsymbol{\beta}_j\|_2}\right)$ 表示块对角矩阵，每个对角块为 $\frac{\sqrt{d_j}}{\|\boldsymbol{\beta}_j\|_2}\boldsymbol{I}_{d_j\times d_j}$，$\boldsymbol{I}_{d_j\times d_j}$ 表示 $d_j\times d_j$ 阶的单位矩阵.

Wei 等人[28]给出的关于组套索的设计矩阵的稀疏 Riesz 条件为：已知 $\widetilde{A}\subseteq\{1,\cdots,J\}$，$|\widetilde{A}|=\Sigma_{j\in\widetilde{A}}j$ 表示集合 $\widetilde{A}$ 中元素的个数，$\boldsymbol{X}_{\widetilde{A}}=(\boldsymbol{X}_j,j\in\widetilde{A})$ 表示由 $|\widetilde{A}|$ 个子设计矩阵组成的新矩阵，令 $\Sigma_{\widetilde{A}\widetilde{A}}=\frac{\boldsymbol{X}_{\widetilde{A}}^{\mathrm{T}}\boldsymbol{X}_{\widetilde{A}}}{N}$，$q^*=|A|$，若对于任意的 $\widetilde{A}$ 和 $\widetilde{\boldsymbol{V}}\in\boldsymbol{R}^{\Sigma_{j\in\widetilde{A}}d_j}$ 都有下式成立：

$$c_*\leq\frac{\|\boldsymbol{X}_{\widetilde{A}}\widetilde{\boldsymbol{V}}\|_2^2}{N\|\widetilde{\boldsymbol{V}}\|_2^2}\leq c^* \tag{35}$$

则称设计矩阵 $\boldsymbol{X}$ 满足关于给定的常数 $q^*$ 和 $0<c_*<c^*<\infty$ 的稀疏 Riesz 条件. 显然,稀疏 Riesz 条件的作用为规定了矩阵变量值的上下界. Wei 等人在组套索的参数估计一致性方面进行了研究,他们给出了组套索在满足稀疏 Riesz 条件和其他一些条件时参数估计值的 $L_2$ 范数误差 $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2$ 的上界.

Lounici 等人[75]给出的关于组套索的设计矩阵的限制特征值条件为:已知 $\tilde{S}\subseteq\{1,2,\cdots,J\}$,$|\tilde{S}|$ 表示集合 $\tilde{S}$ 中元素的个数,$\tilde{S}^C$ 表示 $\tilde{S}$ 的补集,正整数常数 $1\leq s^*\leq J$,若存在一个正的 $c^*$ 使得如下最优化问题的最优值大于 0

$$\min\|\boldsymbol{X}\Delta\|_2/\sqrt{N}\|\Delta_{\tilde{S}}\| \tag{36a}$$

$$\text{s.t.}\quad |\tilde{S}|\leq s^* \tag{36b}$$

$$\Delta\in\boldsymbol{R}^p,\Delta\neq\boldsymbol{0} \tag{36c}$$

$$\sum_{j\in\tilde{S}^C}\|\Delta_j\|\leq c^*\sum_{j\in\tilde{S}}\|\Delta_j\| \tag{36d}$$

则称满足限制特征值条件.

## 9 结构稀疏模型小结

结构稀疏模型按实现的稀疏化效果可分为组结构稀疏模型、双层结构稀疏模型、树结构稀疏模型和图结构稀疏模型. 组结构稀疏模型中,组套索忽视了变量组之间重叠的情况,重叠组套索是在组套索基础上考虑了分组之间重叠的情况而提出的;组套索只是基于线性回归模型的,逻辑斯蒂组套索将组套索从线性回归模型推广到逻辑斯蒂回归模型;组套索不具有变量选择一致性,针对组套索这一缺点而提出的自适应组套索、L2 范数组 SCAD 模型、L2 范数组 MC 模型和 L2 范数组桥模型与组套索相比往往在变量选择一致性和变量组选择一致性方面表现更佳. 在双层结构稀疏模型中,稀疏组套索可实现分组水平上和分组内变量水平上双层的稀疏化结构,但稀疏组套索在变量选择一致性和变量组选择一致性方面表现很差,而 L1 范数组 SCAD 模型、L1 范数组 MC 模型、L1 范数组桥模型和指数组套索内部采用了 L1 范数罚而外部采用了非凸罚,因此其与稀疏组套索相比在变量组选择一致性上往往表现更佳;复合组 MC 模型和复合组桥模型内部与外部均采用了非凸罚,其相对于前述几种内部采用 L1 范数罚的双层稀疏模型来说在组内变量选择的一致性上往往表现更佳. 树结构稀疏模型中,树组套索实现的稀疏效果为解释变量的树结构稀疏化,而多输出树组套索实现的稀疏效果为响应变量的树结构稀疏化. 图结构稀疏模型中,图套索只可实现图的 precision 矩阵中元素稀疏化,而 $L_{\infty,1}$ 范数组结构图套索和 $L_{2,1}$ 范数组结构图套索可实现图的 precision 矩阵中元素的组结构稀疏化. 表 1 对各结构稀疏模型的结构稀疏特性,算法复杂性和不同模型间的联系进行了总结. 表 2 中,$N$ 是样本个数,$P$ 是样例维数,$\max\{d_1,\cdots,d_{|G|}\}$ 为组内模型向量的最大维数,$|G|$ 为组集合的势,$T$ 为迭代次数,$K$ 是输出变量 $\boldsymbol{Y}$ 的维数,$\epsilon$ 是预设的准确率,$J$ 树中的顶点个数,$E$ 为树中的边集.

**表 2 各结构稀疏模型对比**

| 模型 | 结构稀疏性 | 算法复杂性 | 不同模型间的联系 |
|---|---|---|---|
| 套索[75] | 非结构化稀疏性 | $O(NP\min\{N,P\})$ | 利用正则化方法进行变量选择和稀疏化的开山之作 |
| 组套索[76] | 组稀疏性 | $O(P+\max\{d_1,\cdots,d_{\|G\|}\}In\|G\|)$ | 将 Lasso 的非结构化稀疏性推广为组稀疏性 |
| 自适应组套索[77] | 组稀疏性 | $O(TNP\min\{N,P\})$ | 进一步改善了 Group Lasso 的估计准确性 |
| 重叠组套索[78] | 重叠组稀疏性 | $O(P\|G\|)$ | 将 Group Lasso 的组稀疏性推广为重叠组稀疏性 |
| 逻辑斯蒂组套索[79] | 组稀疏性 | $O(2NP^2)$ | 将 Group SCAD 和 MCP 罚推广到 logistic 模型中 |
| $L_2$ Group SCAD[80] | 组稀疏性 | $O(NP^2)$ | 进一步改善了 Group Lasso 的估计准确性 |
| $L_2$ Group MC[81] | 组稀疏性 | $O\left(\frac{1}{\sqrt{\epsilon}}(NP+NK)\right)$ | 进一步改善了 Group Lasso 的估计准确性 |
| $L_2$ Group Bridge[82] | 组稀疏性 | $O(TNP^2)O(NP^2)$ | 进一步改善了 Group Lasso 的估计准确性 |
| 稀疏组套索[83] | 双层稀疏性 | $O(1/\epsilon)$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| 指数组套索[84] | 双层稀疏性 | $O(2NP^2\|G\|)$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| $L_1$ Group SCAD[85] | 双层稀疏性 | $O(NP^2\|G\|)$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| $L_1$ Group MC[86] | 双层稀疏性 | $O(2NP^2\|G\|)$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| $L_1$ Group Bridge[87] | 双层稀疏性 | $O(TNP\min\{N,P\}\|G\|)$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| 复合组桥[88] | 双层稀疏性 | $O(2T\|G\|NP\min\{N,P\})$ | 在 Group Lasso 基础上进一步实现双层稀疏性 |
| 树组套索[89] | 树组稀疏性 | $O\left(\frac{1}{\epsilon}(J^2+\|E\|)\right)$ | 将 Group Lasso 的组稀疏性推广为树组稀疏性 |
| 多任务树组套索[90] | 树组稀疏性 | $O\left(\frac{1}{\epsilon}(J^2K+J\|E\|)\right)$ | 将结构稀疏模型向多任务学习进行拓展 |
| 图套索[91] | 图稀疏性 | $O(\max\{d_1,\cdots,d_{\|G\|}\}+\|G\|\log P)$ | 将稀疏思想应用到概率图模型中,简化概率图模型的学习 |

## 10　优化算法

结构稀疏模型优化算法首先使用控制-受控不等式(Majority-Minority，MM)技术[92]，Nesterov 双目标函数近似方法，或一阶泰勒展开和二阶泰勒展开技术求得原来包含非可微、非凸和不可分离变量的结构稀疏模型目标函数的可微、凸和可分离变量的近似目标函数，然后对结构稀疏模型的近似目标函数使用最小角回归算法、组最小角回归算法[3](Group Least Angle Regression，Group LARS)、块坐标下降算法(block coordinate descent algorithm)[3]、分块坐标梯度下降算法[93](block coordinate gradient descent algorithm)、局部坐标下降算法[94](local coordinate descent algorithm)、谱投影梯度法[7](Spectral Projected Gradient algorithm)、主动集算法(active set algrithm)和交替方向乘子算法(Alternating Direction Method of Multipliers，ADMM)等算法求解最优化解. 可使用 Cp 准则、AIC 准则、BIC 准则和 GCV 准则计算结构稀疏模型的近似目标函数中的权衡参数 $\lambda$，也可以在 $\lambda$ 的取值区间上，选择有限个离散值，使用交叉校验方法确定权衡参数 $\lambda$ 的值，还可以使用解路径方法，通过理论分析给出 $\beta(\lambda)$ 的解路径表达式，找出全局最优值.

(1) 变分不等式. 利用变分不等式可将不光滑的函数等价地转化为光滑的函数，但该方法会引入一个辅助变量，故接下来往往利用轮换方向乘子法继续求解. 例如，在多输出树组套索模型中利用柯西-施瓦兹不等式将非光滑的问题转化为光滑问题然后利用轮换方向乘子法求解.

(2) Nesterov 技巧. 假定 $f=g+h$ 中假定 $g$ 连续可微，而 $h$ 不可微. Nesterov 把损失函数加正则化项双目标最优化问题 $f=g+h$ 中的可微函数 $g$ 用近似函数替代，然后在近似替代后的 $\hat{f}$ 函数上求解最优化问题. 假定 $g$ 的 lipchitz 梯度为 $L$

$$\|g(\beta)-g(\beta')\|_2\leq L\|\beta-\beta'\|_2,\ \forall\beta,\beta'\in\boldsymbol{R}^P \quad (37)$$

使用固定步长 $s^t=s\in(0,1/L]$，则对任意梯度下降迭代次数 $t=1,2,\cdots$，有

$$\hat{f}(\beta)-\hat{f}(\beta^*)\leq\frac{C}{t+1}\|\beta^t-\beta^*\|_2 \quad (38)$$

这里 $C$ 为常数，$\beta^*$ 为最优解，所以在梯度下降迭代过程中，$f(\beta)$ 以 $O(1/t)$ 子线性收敛，实际迭代过程 $L$ 不知时，可使用 Armijo 法则选择步长，可产生同样的收敛率.

Nesterov 提出的方法还包括引入对偶范数，同时将原来变量块不可分离的问题转化成了变量块可分离的问题，经过预处理后的优化问题可以采用梯度法等诸多传统凸优化方法进行求解，该方法非常适用于对重叠组套索和树组套索这类不光滑且变量块不可分离的结构稀疏模型的目标函数进行预处理[92].

(3) 对偶范数和对偶函数. 该预处理手法主要用于解决变量块不可分离的问题，利用引入的对偶变量将原变量块不可分离的问题转化为对偶变量块可分离的问题. 例如，重叠组套索模型和树组套索模型的目标函数均不光滑且变量块不可分离，可经过该方法预处理后得到不光滑但变量块可分离的目标函数，然后利用块坐标下降算法求解.

(4) 局部近似. 该预处理方法分为局部二次近似和局部线性近似两种，主要应用于组 SCAD 模型、组 MC 模型和组桥模型最优化问题，局部近似使用控制-受控不等式(Majority-Minority，MM)技术，把结构稀疏化目标函数替换为二次函数，解决非凸非光滑的 SCAD 罚、桥罚和 MC 罚在进行优化时的困难，但由于涉及海森矩阵的重复求逆问题因而计算复杂度很大，局部二次近似方法还需要设定一个初始解，在初始解处开始迭代，该初始解的设定好坏在很大程度上影响了算法收敛的快慢，并且该方法为了避免数值不稳定问题需要设置一个阈值参数 $\varepsilon_0$ 使得变量的模型向量分量小于该参数时则直接将该变量的模型向量分量置零，因此该方法另一个缺点是其类似于后向变量消除法有丢弃变量不可再恢复的缺点，一旦某变量的模型向量分量被置零，该变量将永远不再会出现在最后求得的模型中，而且阈值参数 $\varepsilon_0$ 的值被设置的大小会影响解的稀疏性和收敛速度，而究竟设置该阈值参数为多大也没有统一的标准. 局部线性近似将原目标函数近似为一次函数，该方法不存在数值不稳定问题，因此不必设置阈值参数，避免了局部二次近似的上述种种问题，但局部线性近似算法性能对初始值的设置比较敏感.

(5) (组) 最小角回归算法. 套索提出后缺乏高效的求解算法，因此没有立刻广泛流行，直到最小角回归算法被提出后，套索等基于正则化方法的稀疏模型受到越来越多的关注，可见最小角回归算法的重要地位. 组套索被提出后，最小角回归算法又被推广为组最小角回归算法，但最小角回归算法和组最小角回归算法均只适用于解路径为分段线性的稀疏模型，这是因为只有稀疏模型的解路径为分段线性

时,由最小角回归算法和组最小角回归算法所求得的解路径才是稀疏模型的解路径.例如,对于组套索模型来说应用组最小角回归算法求解的条件为设计矩阵为正交矩阵,因为只有设计矩阵为正交矩阵时组套索模型的解路径才是分段线性的.组最小角回归算法与最小角回归算法均由于解路径为分段线性函数,可以求解整个解路径,发现最优权衡参数 $\lambda$ 下的稀疏解.

(6)(块)坐标下降算法.块坐标下降算法的前身是坐标下降算法,该算法使用分而治之的策略,把多变量的优化问题,分解为子多变量迭代优化问题,问题的规模变小,靠时间复杂性增加,减少优化问题的复杂性.在迭代过程,任意选择多变量 $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_P)^{\mathrm{T}}\in\boldsymbol{R}^P$ 的一部分变量 $\boldsymbol{\beta}^{(t)}=(\boldsymbol{\beta}_1^{(t)},\cdots,\boldsymbol{\beta}_i^{(t)})^{\mathrm{T}}\in R^i$ 作为要求解的未知变量,$\boldsymbol{\beta}$ 的其余的变量任意置初值,待求解出 $\boldsymbol{\beta}^{(t)}$ 后此子向量作为已知值,再在 $\boldsymbol{\beta}$ 中选择除 $\boldsymbol{\beta}^{(t)}$ 外其余的变量块作为未知值,求解优化问题,此过程反复执行.(块)坐标下降算法只能适用于块可分离的组套索模型.函数变量块可分离性指的是函数可以被写作不同坐标块(坐标向量)$\boldsymbol{\beta}_j$ 对应的函数 $f_j(\boldsymbol{\beta}_j)$ 之间的独立相加关系

$$f(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_J)=\sum_{j=1}^{J}f_j(\boldsymbol{\beta}_j) \tag{39}$$

且不同坐标块(坐标向量)$\boldsymbol{\beta}_j$ 之间不包含共同的自变量.例如,$L_{2,1}$ 组套索模型、自适应组套索模型、混合组套索模型和逻辑斯蒂组套索模型的目标函数均是变量块可分离的,可以直接应用块坐标下降算法对其求解,而重叠组套索模型的目标函数中不同组对应的子模型向量之间存在共同的自变量,因此是变量块不可分离的,所以不能利用块坐标下降算法求解重叠组套索模型.与重叠组套索模型类似,树组套索模型和多输出树组套索模型的目标函数中不同组对应的子模型向量之间也存在共同的自变量,也是变量块不可分离的,因此也不能利用块坐标下降算法直接对其进行求解.某些非凸组稀疏模型的求解也可以利用块坐标下降算法,但要求该模型具有关于单变量的显式解和某些使得目标函数为凸函数的附加条件,例如组桥模型不具有关于单变量的显式解,故不能利用块坐标下降算法求解,$L_2$ 组 SCAD 和 $L_2$ 组 MC 模型具有关于单变量的显式解,可以利用块坐标下降算法求解,但还要附加令其整个目标函数为凸函数的条件才行,这是因为 SCAD 罚和 MC 罚为非凸罚,在某些附加条件下其整个目标函数才是凸的.

(7)局部坐标下降算法.由于基于层次罚的双层稀疏模型不具有关于单变量的解析解,因此不能直接利用坐标下降算法求解.为了解决这个问题,Huang 等人[18]提出了一种局部坐标下降算法,其思想为利用以一阶泰勒展开式对层次罚函数进行近似,得到其 majorizing 函数,利用 MM(Majorize Minirize)方法求解.局部坐标下降算法与坐标下降算法类似,每次只关于单个变量进行优化,不断循环迭代每个变量直到算法收敛从而得到最终的解.

(8)块坐标梯度下降算法.块坐标梯度下降算法首先求解原目标函数的凸二次近似函数,使用 Armijo 法则线搜索确定步长,使用块坐标下降求解结构稀疏化问题的凸二次近似函数的最优解,故块坐标梯度下降算法可被视为梯度法与块坐标下降算法的结合.块坐标梯度下降算法与块坐标下降算法比较,块坐标梯度下降算法适用于大规模问题求解,方法简单,可并行执行,算法收敛快.例如在逻辑斯蒂组套索模型的求解中,关于第 $j$ 个组的优化问题中逻辑斯蒂组套索模型的解 $\boldsymbol{\beta}_j$ 不具有显式更新公式,因此这一步需要数值迭代计算方法,对于大规模的变量选择问题计算复杂度太大,因此块坐标下降算法在该情形下只适用于中小规模的变量选择问题,而块坐标梯度下降算法在该情形下表现更好.

(9)活动集算法.活动集算法一般用于大规模复杂问题的求解,利用最优性条件如 KKT 条件把大规模复杂问题分解为一系列简单子问题的求解,一般是用时间复杂性增高换取空间复杂性的降低,但也可能由于算法迭代过程中出现满足最优性条件提前结束,从而实现同时降低算法的时空复杂性,该算法目前主要用于对 $L_{2,1}$ 组套索模型和 $L_{\infty,1}$ 组套索模型的求解.虽然该算法将大规模的复杂问题转化为一系列简单子问题的求解,但该算法是否高效仍然依赖于子问题的求解是否高效,例如应用于求解 $L_{2,1}$ 组套索模型和 $L_{\infty,1}$ 组套索模型时用于求解子问题的投影梯度步骤计算十分关键[67],很大程度上决定了活动集算法的高效与否.

(10)交替方向乘子算法.由于在第一阶段对结构稀疏化模型进行了近似变换,替换后的目标函数往往具有双变量目标函数约束凸优化形式

$$\begin{aligned} f=&\mathop{\arg\min}_{\boldsymbol{\beta}\in\boldsymbol{R}^P,\theta\in R^M} g(\boldsymbol{\beta})+h(\theta) \\ &\text{s.t. } \boldsymbol{A}\boldsymbol{\beta}+\boldsymbol{B}\theta=\mathrm{c} \end{aligned} \tag{40}$$

这里 $g(\boldsymbol{\beta}):\boldsymbol{R}^P\to R$,$h(\theta):R^M\to R$ 为凸函数,$\boldsymbol{A}\in\boldsymbol{R}^{M\times D}$,$\boldsymbol{B}\in\boldsymbol{R}^{M\times D}$ 为已知矩阵,c 为常数.为了求解上

述约束优化问题，引入拉格朗日乘子 $\mu\in R^D$，得到增广拉格朗日对偶问题

$$L(\boldsymbol{\beta},\theta,\mu)=g(\boldsymbol{\beta})+h(\theta)+\langle\mu,\boldsymbol{A\beta}+\boldsymbol{B}\theta-\mathrm{c}\rangle+\frac{\rho}{2}\|\boldsymbol{A\beta}+\boldsymbol{B}\theta-\mathrm{c}\|_2^2 \tag{41}$$

$\rho>0$ 为固定常数，$\frac{\rho}{2}\|\boldsymbol{A\beta}+\boldsymbol{B}\theta-\mathrm{c}\|_2^2$ 为平滑项. 交替方向乘子算法迭代更新公式为

$$\beta^{(t)}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P} L(\boldsymbol{\beta},\theta^{(t-1)},\mu^{(t-1)}) \tag{42a}$$

$$\theta^{(t)}=\arg\min_{\theta\in R^M} L(\beta^{(t)},\theta,\mu^{(t-1)}) \tag{42b}$$

$$\mu^{(t)}=\mu^{(t-1)}+\rho(\boldsymbol{A}\beta^{(t)}+\boldsymbol{B}\theta^{(t)}-\mathrm{c}) \tag{42c}$$

对于拉格朗日乘子 $\mu\in R^D$，迭代过程 $t=1,2,\cdots$ 为对偶上升过程.

由于目标函数对变量 $\boldsymbol{\beta}$ 和 $\theta$ 是分离的，$g(\boldsymbol{\beta}):\boldsymbol{R}^P\to R$ 只依赖于 $\boldsymbol{\beta}$，$h(\theta):R^M\to R$ 只依赖于 $\theta$. 交替方向乘子算法适宜于求解非可微约束下的凸问题，可以把数据样例分为成多个块，在每个数据块上分别运用交替方向乘子算法求解，从而实现大规模问题分解为多个小问题；还可以对高维问题，每个样例的各维变量划分为多个块，以坐标块方式求解.

轮换方向乘子法适用于分布式和并行算法求解，可用来求解各种结构稀疏模型.

(11) 谱投影梯度法. 谱投影梯度法是对投影梯度算法的改进，主要用于克服后者收敛速度慢的缺点. 投影梯度算法存在两方面的问题：一是每次选择最速下降方向会导致收敛速度变慢；二是投影步骤的计算复杂度过高. 谱投影梯度法使用非单调线搜索技术确定步长，算法迭代过程目标函数可能不总是下降的，并且需要结合谱梯度法的 Barzilai-Borwein 步长来选择谱投影梯度法在迭代过程中的步长. 谱投影梯度法适用于投影步骤计算高效的情形，因此投影步骤的计算方法非常关键，当前该方法主要用于对于 $L_{2,1}$ 组套索模型和 $L_{\infty,1}$ 组套索模型组套索模型的求解[95].

(12) 小结与分析. 不同结构稀疏模型所用来求解的方法(预处理方法与优化算法)不同，不同的算法也有各自的适用条件，因此在面对一种新的组稀疏模型时，往往需要从该组稀疏模型的目标函数的特点出发，根据其目标函数是否平滑、是否变量块可分离和是否为凸函数等方面进行全盘考虑从而确定最佳的求解方法. 各预处理方法与求解算法的优点与不足如表 3 所示，文献中已有的对各种组稀疏模型求解算法如表 4 所示. 陶等人[96]综述了稀疏模型的算法，读者可参考之.

**表 3　各预处理方法和算法的优点与不足**

| 算法 | 优点与不足 |
|---|---|
| 变分不等式 | 通过引入新的变量，将非光滑函数等价地转化为光滑函数，适当选择新增加的变量的值可使转化后的问题的最优解是原问题的解，缺点是优化的自变量数增多，增加了问题的求解复杂性. |
| Nesterov 平滑近似技巧 | 同时解决了不平滑与变量块不可分离问题，缺点是要求原问题满足强凸条件，求得的最优解只是原问题的近似解. |
| 对偶范数和对偶函数 | 通过引入对偶变量，求解对偶变量的凸最优化问题，可求解变量块不可分离原问题，缺点是算法复杂性高，无法用于大规模结构稀疏化问题求解. |
| 局部二次近似 | 解决非凸与非平滑问题，但导致计算复杂、丢弃变量不可再恢复、依赖于阈值参数且数值不稳定，算法性能的好坏很大程度上依赖于初始解的设定. |
| 局部线性近似 | 克服局部二次近似方法的计算复杂、丢弃变量不可再恢复、依赖于阈值参数和数值不稳定等缺点，但算法性能的好坏很大程度上依赖于初始解的设定. |
| 组最小角回归 | 可得到整个解路径，但要求问题的解路径是分段线性的，大多数问题不满足此条件，应用范围有限. |
| 块坐标下降 | 有效地将复杂优化问题转化为容易求解的简单问题，但要求目标函数变量块可分离或具有关于子变量的显式解，要求子问题的求解复杂性尽可能低. |
| 局部坐标下降算法 | MM 算法与坐标下降算法的结合. |
| 块坐标梯度下降 | 梯度法与块坐标下降算法的结合，简单且具有高度并行化的特点，并且收敛性较强，尤其在块坐标下降算法无显式解时体现出其优势，适用于求解大规模的优化问题. |
| 活动集算法 | 有效地将大规模复杂问题分解为一系列活动集变量上简单子问题的求解，但算法复杂性仍然依赖于子问题求解算法是否高效，且活动变量选择需要前向和反向选择过程，时间复杂性高. |
| 交替方向乘子算法 | 适宜于求解非可微约束下的凸问题，可对数据样例分为成多个块或每个样例的各维变量划分为多个块，可用于高维大规模问题并行分布式求解. 缺点是算法只适用于凸问题，且引入了新的乘子迭代过程，要求分解子问题的求解复杂性低. |
| 谱投影梯度法 | 克服了投影梯度收敛速度慢的缺点，但整个算法的有效性依赖于投影步骤的计算复杂度. |

**表 4　文献中对各种组套索模型的求解算法**

| 模型 | 求解算法 |
|---|---|
| 组套索模型 | 组最小角回归算法，块坐标下降算法，轮换方向乘子法，活动集算法，谱投影梯度法 |
| $L_{\infty,1}$ 组套索模型 | 活动集算法，谱投影梯度法 |
| 自适应组套索模型 | 块坐标下降算法 |
| $L_2$ 组 SCAD 模型 | 块坐标下降算法 |
| $L_2$ 组 MC 模型 | 块坐标下降算法 |
| $L_2$ 组桥模型 | 局部坐标下降算法 |
| 稀疏组套索模型 | 坐标下算法＋块坐标下降算法 |
| 逻辑斯蒂组套索模型 | 块坐标下降算法，块坐标梯度下降算法 |
| 基于层次罚的双层结构稀疏模型 | 局部坐标下降算法 |
| 重叠组套索模型 | 轮换方向乘子法，Nesterov 平滑近似技巧＋块坐标下降 |
| 树组套索模型 | 轮换方向乘子法，Nesterov 平滑近似技巧＋块坐标下降 |
| 多输出树组套索模型 | 变分不等式＋轮换方向乘子法，Nesterov 平滑近似技巧＋块坐标下降 |
| 图套索 | 块坐标下降算法 |

## 11 结构稀疏模型的应用

总体来说,结构稀疏模型在模型预测和特征选择上取得了很大的成功.下面以结构稀疏模型在生物信息学和医药学、在时变和空间暂态数据上的应用,在文本处理中的应用,在图像检索和图像理解的应用和在压缩传感和图像重构中等典型应用予以介绍.

### 11.1 在生物信息学和医药学中的应用

在基因网逆向工程和发现与脑功能关联的模式等任务中,目标为重构感兴趣的变量之间的依赖关系,找出与所感兴趣疾病最相关的基因或基因组和脑区域.

文献[44]提出树结构稀疏罚模型应用 DNA CGH(Comparative Genomic Hybridization)微阵列数据诊断癌症疾病.

在神经科学中,让受试者说一段与某个主题有关的话,或者看一些特定的图片,记录下同一脑区域功能磁共振(functional Magnetic Resonance Imaging,fMRI)三维像素(voxels)[93],提取 1700 到 2200 的三维像素子集作为输入特征,构造稀疏高斯马尔科夫网预测人的心理状态,使用 SINCO 算法(sparse Gaussian MRF models learned by an algorithm),完成了 5%的错误率,显著好于 SVM 预测结果[97].

文献[98,99]给出另一个使用稀疏马尔科夫网对心理状态预测的成功应用例子.这里,目标为发现局部脑功能异常的精神分裂症精神疾病患者最相关的脑区域 fMRI 三维像素特征,即统计生物学标记物,即生物学家认为从 fMRI 得到的精神分裂症和健康受试者执行简单的听觉任务的三维像素,从整个脑部获取的 fMRI 三维像素特征用于构造脑功能相互作用网络,该网络的拓扑特征包含了丰富的信息,实验表明,稀疏马尔科夫网能够准确地判别两组受试者,得到了 86%的正确率.稀疏马尔科夫网也用于阿尔茨海默氏症[100]和吸毒引起的脑紊乱研究[101].稀疏马尔科夫网模型能够辨识与相应疾病引起的网络结构发生的改变,文献[102]进一步引入有节点和边选择的稀疏马尔科夫网,这里分组由同一节点变量的相邻边组成,同一组边置为 0 时,去除相应的节点,该模型能极大地改进马尔科夫网的解释性,特别是马尔科夫网中节点上千个时,尤为明显.

### 11.2 在时变和空间暂态数据上的应用

在酵母细胞周期微阵列时间过程基因表达数据上,应用组 SCAD 回归模型,能够有效地辨识变系数生物过程基因调控转录因子,能够正确地发现已知的与细胞周期过程有关的 21 个转录因子中的 19 个,而且能辨识其他与细胞周期过程有关的周期转录效果的 52 个转录因子.实验结论为组 SCAD 回归模型非常适合估计和选择时变过程有关变量,能够发现随时间变化的转录因子[103].

气候预测数据具有高维和空间暂态特性,需要使用具有变量选择功能的模型,文献[104]使用海洋气候数据构造陆地气候预测模型,采用树结构稀疏组范数作为正则化项回归模型,论文证明了统计估计一致性,实验表明优于以前的预测性能,而且模型是气候学上可解释的.

### 11.3 在文本处理中的应用

Web 和社交媒体数据上大量存在短文本数据(Short Text Classification,STC),短文本数据相较于普通文本(Text Classification,TC)具有特征稀疏性,以前的适用于一般文本的模型不再适用,存在的处理 STC 的方法是附加新的信息或将外来语料库添加到 TC 里,如果选择外来内容数量和性质不适当,会使预测得性能恶化,且不一定存在合适的外来语料库,获取外来语料库也需要人力,文献引入凸包顶点选择减少字典冗余和相关系数构造 STC 结构稀疏模型,在 5 个 STC 数据上分类结果好于以前最好的使用外部语料库的分类效果[105].

使用隐主题分布先验把隐主题作为稀疏退避树(Sparse Backoff Tree,SBT)的叶子节点,利用稀疏树结构上的 collapsed 抽样方法,基于 SBT 文本模型能够在上百万主题上有效地推理主题[106].

主题模型在发现医疗,金融和计算机视觉中的隐结构很有用,但是,即使使用稀疏技术,还是很难发现可解释的主题,使用本体单词之间的关系上的图稀疏 LDA,在生物医疗数据上,建立层次主题模型,得到更少的可解释的隐概念-单词主题[107].

多文本总结能够节省阅读时间,很快地发现文本中的主要的内容.基于数据重构和语句解噪模型,提出两级稀疏表示模型,使用模拟退火算法在基准数据 DUC2006 和 DUC2007 上从多个文本集提取总结,实现重构总结,实现了满足全覆盖、稀疏和多样性的文本总结[108].

### 11.4 在图像检索和图像理解的应用

神经元空间暂态荧光动力可以表示为光场中每一个神经元空间位置的矩阵与每一个神经元随着时间变化的钙离子指示剂中心矩阵的乘积,结构矩阵

因式化神经团钙离子成像，可应用于活体树突状大规模神经元成像数据中的空间重叠分量解混合，以及钙离子指示剂动态变化中神经元脉冲活动的解噪和解卷积，可估计钙离子指示剂动力曲线[109].

在图像检索和图像理解中的图像自动标注问题(Automatic Image Annotation，AIA)，文献[110]研究了结构化视觉图像特征选择和层次相关结构多标注问题，提出同时实现输入输出结构组稀疏图像标注正则化回归模型，提升了图像标注的性能.考虑到输入高维异构特征如颜色、纹理以及形状等，不同的特征具有不同的概念识别分辨力，文献[111]提出的同时考虑输入输出结构组稀疏图像标注正则化回归模型的结构特征选择方法 Bi-MtBGS(Bilayer regression model for Multilabel Boosting by the selection of heterogeneous features with structural Grouping Sparsity)能实现组内和组级特征选择，输出类标签层次相关结构用稀疏树结构表示，提出的回归模型具有两层结构，第一层回归模型实现每一个标签的判别特征选择，第二层回归为多标签集成方法，在基准数据上的实验结果显示模型增加了图像理解的可解释性.

计算机视觉视频理解人类行为识别中，假定视频分解为关键词分量，局部图像特征作为词汇表的视觉单词组成袋子模型(Bag of visual Words，BoW)，采用人类行为中的暂态结构组稀疏码的系数构造直方图作为 BoW，实时采集视觉单词结构稀疏系数作为人类行为分量的几何特征，文献[112]提出的方法量化误差小，暂态结构减少了模型参数和存储空间复杂性，在基于人类行为识别数据集 KTH、Weismann 和 UCF-sports and UCF50 上的实验结果表明比先前的方法要好.

高光谱图像研究中，文献[113]考虑构造属于同一类标签的类内聚类簇样例保持图结构低维投影，不需要像传统线性判别分析那样预先知道已标识样例个数，提出的稀疏图判别分析可以用广义特征值方法求解，实验显示算法同时实现降维和判别分析.

### 11.5　在压缩传感和图像重构中的应用

信号重构中需要发现所要查询信号的最近树稀疏信号，即树投影问题，长度为 $n$ 的信号投影到稀疏参数为 $k$ 的树结构，信号的取值区间为 $r$，原来最优的算法的复杂性为 $O(nk)$，提出的近似算法具有 $O(n\log kr)$存储复杂性，算法可应用于大规模高维树稀疏压缩传感问题[114].

压缩传感研究中，某个基函数上的未知信号的稀疏表示特性可用作信号重构时的先验知识.基于结构稀疏表示模型重构算法除了利用稀疏先验信息外还使用了未知信号的特殊结构信息，分段平滑信号的小波变换上常用的结构为稀疏树结构，文献[115]提出一种利用稀疏树先验信息压缩传感模型，使用基于树的正交匹配追踪算法求解压缩传感问题.

文献[116]提出结构稀疏信号的支集估计自适应压缩传感算法，该算法对观测数据噪声鲁棒，文献同时给出了自适应压缩传感的估计一致性的上界和下界.

文献[117]考虑盲聚类结构稀疏重构问题，实现了基于模型的非线性滤波压缩传感，提出的模型假定稀疏信号非 0 分量组成块结构，非 0 块数，维数和位置均未知，引入确定性非随机近邻非凸非可分罚，用 L1 迭代赋权方法，提出的模型不需要稀疏模式任何信息，实验结果显示压缩传感性能优于以前的研究结果.

文献[118]利用多视图构造语义一致的图像特征表示，提出组稀疏多视图块对齐框架，使得各视图语义相互补充，利用投影矩阵 $L_2$ 范数行稀疏性同时实现特征提取和特征选择功能，实验显示提高了图像分类性能.

## 12　实验分析

### 12.1　40 个变量实验比较

本节选取套索、组套索和复合组 MC 模型来比较其实现的(结构)稀疏化效果.组套索代表了 $L_2$ 组 SCAD 模型和 $L_2$ 组 MC 模型等一大类组稀疏模型，而复合组 MC 模型代表了稀疏组套索、$L_1$ 范数组 MC 模型、$L_1$ 范数组 SCAD 模型以及复合组桥模型等一大类双层结构稀疏模型.本实验中，使用的 R 语言包分别为 cran 网站上的 glmnet 包(https://cran.r-project.org/web/packages/glmnet/index.html)、gglasso 包(下载网址为 https://cran.r-project.org/web/packages/gglasso/index.html)和 grpreg 包(下载网址为 https://cran.r-project.org/web/packages/grpreg/index.html).

为了验证复合组 MC 模型的双层变量结构稀疏化效果，生成一基于线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\varepsilon$ 的人工数据集，该人工数据集中共包含 40 个变量，各变量的分组情况及其真实系数值如下：

$$\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_7)^{\mathrm{T}}$$

并且上述向量中各个分量分别为

$$\boldsymbol{\beta}_1=(1,1,0,0,0),$$

$$\boldsymbol{\beta}_2=(2,2,2,2,0),$$

$$\boldsymbol{\beta}_3=(0,0,0,0,0),$$

$$\boldsymbol{\beta}_4=(0,0,0,0,0),$$

$$\boldsymbol{\beta}_5=(0,0,0,0,0),$$

$$\boldsymbol{\beta}_6=(0,0,3,3,3),$$

$$\boldsymbol{\beta}_7=(0.26,0,0,0,0,0,0,0,0,0).$$

实验结果如表 5、表 6、表 7 和表 8 所示，表中未列出的变量均为实验结果中未被选中的变量.

**表 5 套索的实验结果**

| 变量 | 系数 |
|---|---|
| $V1$ | 0.74 |
| $V2$ | 0.82 |
| $V3$ | −0.03 |
| $V11$ | 1.92 |
| $V12$ | 1.81 |
| $V13$ | 1.92 |
| $V14$ | 2.00 |
| $V24$ | 0.02 |
| $V28$ | 2.88 |
| $V29$ | 3.01 |
| $V30$ | 2.91 |
| $V31$ | 0.13 |
| $V39$ | −0.06 |

**表 6 组套索的实验结果 1**

| 变量 | 系数 |
|---|---|
| $V1$ | 0.25 |
| $V2$ | 0.49 |
| $V3$ | −0.18 |
| $V4$ | 0.03 |
| $V5$ | 0.009 |
| $V11$ | 1.95 |
| $V12$ | 1.63 |
| $V13$ | 1.97 |
| $V14$ | 1.86 |
| $V15$ | −0.27 |
| $V26$ | −0.06 |
| $V27$ | −0.01 |
| $V28$ | 2.90 |
| $V29$ | 2.72 |
| $V30$ | 2.75 |

**表 7 组套索的实验结果 2**

| 变量 | 系数 |
|---|---|
| $V31$ | 0.002 |
| $V32$ | 0.001 |
| $V33$ | −0.001 |
| $V34$ | −0.0001 |
| $V35$ | −0.003 |
| $V36$ | 0.001 |
| $V37$ | 0.0001 |
| $V38$ | −0.001 |
| $V39$ | −0.004 |
| $V40$ | 0.001 |

**表 8 复合组 MC 模型的实验结果**

| 变量 | 系数 |
|---|---|
| V1 | 0.86 |
| V2 | 0.78 |
| V11 | 2.04 |
| V12 | 1.87 |
| V13 | 2.04 |
| V14 | 2.12 |
| V28 | 2.99 |
| V29 | 3.11 |
| V30 | 3.06 |

对比套索和组套索的实验结果，容易看出套索只得到了普通的非结构稀疏化效果，其变量选择结果没有任何结构化形式. 而组套索却实现了组结构化的稀疏性，其实验结果中位于同一个变量组的变量要么同时被选中要么同时不被选中.

对比套索和复合组 MC 模型的实验结果，容易看出复合组 MC 模型没有选中变量 $V40$，而套索却选中了该变量，复合组 MC 模型没有选中变量 $V40$ 的原因为 $V40$ 与 9 个真实回归系数为 0 的冗余变量位于同一组，大量非重要变量的存在导致该变量组为非重要的变量组，$V40$ 受到这 9 个同组中非重要变量的牵扯，从而没有被选中，也就是说复合组 MC 模型中属于同一个组的 10 个变量没有同时被选中，即实现了组稀疏性.

对比组套索与复合组 MC 模型的实验结果，容易看出 Group Lasso 只实现了变量组选择，被分为同一个变量组的变量的系数同时为零或同时非零，复合组 MC 模型实现了组内的变量选择，同一个分组内的变量系数有的为零有的非零.

综上所述，复合组 MC 模型具有双层结构稀疏化效果，而套索不具有结构稀疏化效果，组套索具有组结构稀疏化效果.

### 12.2 人工数据实验比较

人工数据实验比较分为回归和分类实验. 回归时，如果是预设 3 个变量组成组关系，使用

$$\boldsymbol{Y}=\sum_{j=1}^{J}\left(\frac{2}{3}X_{ik}^{j}-X_{ik}^{j}+\frac{2}{3}X_{ik}^{j}\right)\boldsymbol{\beta}_j+\varepsilon,$$

$$\boldsymbol{\beta}_j=(-1)^j\exp[-(2j-1/20)]$$

产生人工数据，多个变量为一组的生成过程类似. $\varepsilon\sim N(0,\sigma^2)$ 为高斯噪声. 按信噪比 $SNR$(Signal-to-Noise Ratio)

$$SNR=\frac{1}{\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}E\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}\boldsymbol{\beta}$$

值为 1 添加高斯噪声.

分类时对上述输出进行逻辑斯蒂函数变换得到

类标签：

$$p(\boldsymbol{Y}=-1|\boldsymbol{X},\boldsymbol{\beta})=1/[1+(1+\mathrm{e}^{-Y})],$$

$$p(\boldsymbol{Y}=+1|\boldsymbol{X},\boldsymbol{\beta})=1/[1+(1+\mathrm{e}^{Y})].$$

取 $N=500$、$P=200$，在人工数据集上，Group lasso、Group MCP、Group SCAD 和 Group Bridge 采用局部二次近似(Local Quadratic Approximation，LQA)，局部线性近似 (Local Linear Approximation (LLA)，局部坐标下降(Local Coordinate Descent，LCD)算法，对运行时间进行了比较，结果见表 9.

取 $N=500$、$P=200$ 和 $J=20$，对 Group lasso、Group MCP、Group SCAD 和 Group Bridge 的预测性能进行了对比. 特征选择的性能按真正是 0 的组未被选中计算得到虚假的正发现率 $FP$(False Positive)和按真正非 0 的组未被选中计算得到虚假的负发现率 $FN$(False Negatives)，100 次平均实验结果见表 10，表 10 中 $GN$ 为选中的组数，$VN$ 为变变量选择个数. 从表 10 可以看出，Group MCP、Group SCAD 选择变量个数比 Group lasso 少，模型更加稀疏；Group MCP 比 Group SCAD 选择更少的变量，更加稀疏.

**表 9　人工数据各结构稀疏模型运行时间对比**

| 任务 | 模型 | 时间/s |
|---|---|---|
| 回归 | Lasso(glmnet) | 0.03 |
| | Lasso(lars) | 0.42 |
| | Group lasso(LQA) | 3.59 |
| | Group lasso(LCD) | 0.62 |
| | Group MCP(LLA) | 5.13 |
| | Group MCP(LCD) | 0.12 |
| | Group SCAD(LQA) | 0.32 |
| | Group Bridge(LLA) | 7.02 |
| | Group Bridge(LCD) | 0.11 |
| 分类 | Lasso(glmnet) | 0.24 |
| | Lasso(glmpath) | 13.77 |
| | Group lasso(LQA) | 21.78 |
| | Group lasso(LCD) | 1.80 |
| | Group MCP(LLA) | 15.08 |
| | Group MCP(LCD) | 0.47 |
| | Group SCAD(LQA) | 16.34 |
| | Group Bridge(LLA) | 29.77 |
| | Group Bridge(LCD) | 0.67 |

**表 10　人工数据各结构稀疏模型性能对比**

| 任务 | 模型 | 组数 | 组选择 | | | 变量选择 | | |
|---|---|---|---|---|---|---|---|---|
| | | | $GN$ | $FP$ | $FN$ | $VN$ | $FP$ | $FN$ |
| 回归 | Group lasso | 20.0 | 6.8 | 0.8 | 0.4 | 58.5 | 20.7 | 1.2 |
| | Group MCP | 4.5 | 10.9 | 3.0 | 0.1 | 24.6 | 7.5 | 3.9 |
| | Group SCAD | 5.4 | 12.8 | 4.3 | 0.1 | 30.6 | 10.1 | 3.12 |
| | Group Bridge | 8.7 | 5.5 | 0.3 | 0.8 | 19.9 | 5.2 | 4.3 |
| 分类 | Group lasso | 20.0 | 5.9 | 0.2 | 0.2 | 61.9 | 7.3 | 2.4 |
| | Group MCP | 5.2 | 10.8 | 2.6 | 0.0 | 28.4 | 4.7 | 14.3 |
| | Group SCAD | 7.7 | 15.4 | 3.6 | 0.0 | 30.5 | 6.8 | 11.7 |
| | Group Bridge | 10.77 | 5.5 | 0.3 | 0.8 | 23.8 | 2.1 | 14.3 |

### 12.3　实际数据实验比较

按表 11 选择 8 组数据，数据特性如表 11 所示，表 11 中 $N$ 为样本个数，$P$ 为样例维数，$J$ 为样例维数中的分组个数. 由于连续变量不知道分组情况，所以采用 5 阶样条基函数作为加模型拟合原有数据，把样条基函数的展开式的系数作为组变量，从而实现分组.

**表 11　实验数据**

| 数据 | $N$ | $J$ | $P$ |
|---|---|---|---|
| Autompg[119] | 392 | 7 | 35 |
| Bardet[120] | 120 | 200 | 1000 |
| Cardiomypathy[121] | 30 | 6319 | 31595 |
| spectroscopy[122] | 103 | 100 | 500 |
| Breast[123] | 42 | 22283 | 111415 |
| Colon[124] | 62 | 2000 | 10000 |
| Prostate[125] | 102 | 6033 | 30165 |
| Sonar[126] | 208 | 60 | 300 |

使用误差公式

$$ER=(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})^{\mathrm{T}}E\{\boldsymbol{X}_{i}.\boldsymbol{X}_{i}^{\mathrm{T}}\}(\boldsymbol{\beta}-\boldsymbol{\beta}^{*})$$

计算误差，这里，$\boldsymbol{\beta}^{*}$为真实模型向量.

实验时，每组数据划分为 60% 为训练数据，40%为测试数据，每组数据实验执行 100 次，对实验结果求平均后计算得到最终实验比较结果. 各结构稀疏模型在 8 组数据上的运行时间对比如表 12 所示. 表 13 为各结构稀疏模型 4 组数据上回归时的性能对比. 表 14 为各稀疏模型在 4 组分类数据集上的性能.

表 12 各结构稀疏模型运行时间对比 (单位:s)

| 模型 | 回归 | | | | 分类 | | | |
|---|---|---|---|---|---|---|---|---|
| | Autompg | Bardet | Cardiomypathy | spectroscopy | Breast | Colon | Prostate | Sonar |
| Group lasso | 3.14 | 9.96 | 78.23 | 9.37 | 439.76 | 60.42 | 111.75 | 24.55 |
| Group MCP | 4.47 | 14.22 | 111.37 | 14.17 | 626.07 | 86.77 | 159.08 | 78.77 |
| Group SCAD | 0.28 | 0.88 | 6.98 | 0.84 | 39.22 | 5.38 | 9.97 | 2.22 |
| Group Bridge | 6.23 | 19.77 | 155.21 | 18.60 | 872.51 | 119.77 | 221.77 | 48.77 |

表 13 各结构稀疏模型回归性能对比

| 模型 | Autompg | | Bardet | | Cardiomypathy | | spectroscopy | |
|---|---|---|---|---|---|---|---|---|
| | *FN* | *ER* | *FN* | *ER* | *FN* | *ER* | *FN* | *ER* |
| Group lasso | 15.5 | 0.343 | 15.2 | 0.321 | 14.9 | 0.474 | 15.7 | 0.333 |
| Group MCP | 4.4 | 0.255 | 4.5 | 0.311 | 4.3 | 0.384 | 4.6 | 0.312 |
| Group SCAD | 12.5 | 0.272 | 11.8 | 0.329 | 12.3 | 0.429 | 12.7 | 0.429 |
| Group Bridge | 3.2 | 0.244 | 3.1 | 0.322 | 3.3 | 0.422 | 3.5 | 0.325 |

表 14 各结构稀疏模型分类性能对比

| 模型 | Breast | | Colon | | Prostate | | Sonar | |
|---|---|---|---|---|---|---|---|---|
| | *FN* | *ER* | *FN* | *ER* | *FN* | *ER* | *FN* | *ER* |
| Group lasso | 15.6 | 0.388 | 14.8 | 0.161 | 15.4 | 0.381 | 14.6 | 0.229 |
| Group MCP | 4.2 | 0.342 | 4.1 | 0.129 | 4.5 | 0.321 | 4.6 | 0.202 |
| Group SCAD | 12.2 | 0.365 | 12.7 | 0.146 | 11.5 | 0.343 | 12.7 | 0.219 |
| Group Bridge | 3.1 | 0.354 | 3.3 | 0.159 | 3.5 | 0.351 | 3.3 | 0.222 |

# 13 未来研究方向

## 13.1 向其他统计模型拓展结构稀疏模型

将组套索和非凸罚组稀疏模型扩展到 Probit 回归模型、负二项回归模型(negative binomial regression models)、Poisson 回归模型、多项逻辑斯蒂回归模型(multinomial logistic regression model)等广义线性模型、索引模型(Index model)、部分线性模型(Partially Linear Models)、变系数模型(varying coefficient models)等半参数回归模型以及 AR 模型,MA 模型和 ARMA 模型、生存分析中的加速失效时间模型(accelerated failure time model)等情形,极大地丰富组稀疏模型,拓展结构稀疏模型的应用范围.

## 13.2 引入其他非凸罚函数到结构稀疏模型中

$L_0$ 范数作为罚函数可以得到稀疏解,但是 $L_0$ 范数作为罚函数时的不连续性和非凸性使得最优化问题求解很困难. $L_1$ 范数对其进行了放松,也能得到稀疏解,但是 $L_1$ 范数对应的套索往往导致过惩罚模型. 针对 $L_0$ 范数和 $L_1$ 范数的上述缺点,很多类似的非凸罚函数陆续被提出,例如 SCAD 罚函数、MC 罚函数、桥罚函数、capped-$L_1$ 罚函数[127,128]、对数和罚函数(Log-Sum Penalty, LSP)[129]、Geman 罚函数[130](Geman Penalty, GP)以及 $L_q$ 范数($0<q<1$)罚[131]等. $L_1$ 范数、SCAD 罚函数、桥罚函数和 MC 罚函数已经被推广到组稀疏模型中形成组套索、组 SCAD 罚模型、组桥模型和组 MC 罚模型,那么把其他 capped-$L_1$ 罚函数、对数和罚函数、Geman 罚函数以及 $L_q$ 范数($0<q<1$)罚等非凸罚函数推广到变量组选择情形下从而得到相应的结构稀疏模型的问题值得研究.

## 13.3 结构稀疏模型的在线算法

Yang 等人[132]给出了组套索的在线学习算法,这种在线学习算法还适用于稀疏组套索和重叠组套索的求解. 未来有待于研究关于组 SCAD 罚模型、组桥模型和组 MC 罚模型等非凸结构稀疏模型的在线算法,并给出算法的收敛性和错误误差界.

## 13.4 结构稀疏模型的统计性质

当前,很多结构稀疏模型的统计性质研究仍是空白. 另外,最近 van de Geer 等人[133]、Zhang[134]和 Ye 等人[135]均给出了限制特征值条件的变体,这些变体比以往的限制特征值条件要强,他们在这些限制特征值条件的变体下对套索模型的统计性质进行了研究,那么结构稀疏模型在这些更强的限制特征值条件的变体下的一致性如何?

## 13.5 将稀疏化和结构稀疏化思想应用到概率图模型的学习中

将稀疏化与结构稀疏化思想应用到概率图模型的学习中值得探究. 最近,有学者分别将图套索推广

到了部分随机变量不可观的图模型[136-139]、有向无环图模型[140,141]、伊辛模型[142,143]、半参数高斯无向图模型[144]、泊松无向图模型[145]、基于广义线性模型的概率图模型[146]、无标度网络[147]和多任务学习等情形[148,149]，将图套索继续向其他类型的概率图模型等情形进行推广值得探究.

### 13.6　权衡参数选择问题

使用 AIC 和 BIC 准则选择模型的权衡参数时，要求估计误差的协方差矩阵和自由度，K-倍交叉校验方法不需要估计误差的协方差矩阵和自由度，但是计算量较大. 最近，Meinshausen 等人[150]基于重抽样(resample)方法提出选择模型的权衡参数的稳定选择(Stability Selection)方法，这种方法是一种通用的权衡参数选择方法，当然也适用于组稀疏模型，重要的是这种方法不要求对方差或自由度进行估计，未来需要做的工作是利用这种方法选择结构稀疏模型的权衡参数，并将其效果与之前的模型权衡参数选择方法，如 Cp 判据、BIC 判据、AIC 判据、GCV 准则和交叉校验等方法进行比较.

### 13.7　考虑变量概率依赖关系和研究非确定性分组结构稀疏模型

在图套索模型中，变量之间的关系可能依赖于变量的值，例如如果 $x1=1$，那么 $x2$ 与 $x3$ 是非常相关的；如果 $x1=0$，那么 $x2$ 与 $x4$ 是非常相关的. 而且特征变量分组也可以是概率的而非确定性的. 这也是一种可能的进一步研究方向.

### 13.8　同时考虑样例分组结构和样例的变量分组结构关系的结构稀疏模型

结构稀疏模型还可以引入样例级结构依赖关系，特征变量之间的关系也有可能与样例有关. 一个例子是数据是有关用户的各种行为描述的元组，那么描述男性和女性用户各种行为关系的样例可能是不一样的. 所以一种可能的研究方向是将变量分组和训练样本分组(如使用 $k$-近邻、高斯混合模型和狄氏混合模型等无监督学习过程对样例进行分组)结合起来，这样能更充分地利用数据上存在的先验信息，实现更加符合数据本身特性的结构稀疏化模型.

### 13.9　在大数据分析中的应用

在科学、技术、商业、国防、电信、搜索引擎、电子商务、社交网络、多媒体、信号处理、计算机视觉、生物信息学和核磁共振影像学等领域，处理的数据往往为无法装载进内存中的数据，即大数据[151]. 大数据具有以下特性：(1) 样本个数大，样本维数高. 处理一遍整个数据需要很长时间，大数据不能装载进计算机内存储器处理，需要流数据摘要算法或者序列增量及在线学习算法；要处理的数据可能来自于或存储于不同的地方，或者需要把数据分配到多个计算机上用分布式或并行算法处理；(2) 数据特性存在概念漂移问题，不断改变的数据特性需要机器学习算法具有模型自适应学习和不断学习数据新特性的能力，而且数据的概率分布往往为重尾分布，不符合许多常规的统计理论规律，需要特殊的统计手段；(3) 需要开发适用于大规模问题的实时分析和决策的机器学习技术[152,153].

稀疏模型和结构稀疏模型可在消除冗余特征方面起到作用. 可考虑：(1) 研究面向高维流数据的在线结构稀疏模型和算法；(2) 研究并行和分布式稀疏结构化模型和求解算法；(3) 研究具有时变和自适应特征的结构稀疏化模型和求解算法，实现动态的结构稀疏模型；(4) 研究面向重尾分布的贝叶斯结构稀疏化模型和相应的求解算法.

随着与结构稀疏模型相配套的优化算法不断发展，例如(块)坐标下降和 ADMM 算法等的提出与应用[154-157]，稀疏模型在大规模数据中的应用已成为可能，ADMM 算法适用于求解损失函数＋正则化项形式的优化问题，而结构稀疏模型的形式恰恰具有损失函数＋正则化项的统一形式，Boyd 等人[158]对于 ADMM 算法在分布式平台 hadoop 上应用的框架进行了详细讨论. 另外，将坐标下降类型的算法应用到分布式计算平台的研究也逐渐出现.

## 14　结论与展望

本文按照组结构稀疏模型、双层结构稀疏模型、树结构稀疏模型和图结构稀疏模型的顺序系统地综述了结构稀疏模型，并且对其求解算法也进行了介绍. 接着，本文综述了研究结构稀疏模型的参数估计一致性、变量选择一致性和 oracle 性质等统计特性时的一些附加条件：不可表示条件、稀疏 Riesz 条件和限制特征值条件. 另外，还综述了结构稀疏化理论与贝叶斯观点、编码复杂度理论和子模函数之间的联系. 最后，本文揭示了结构稀疏模型未来的研究方向.

结构稀疏模型作为当下主流的建模方法，无论是理论研究，还是求解算法都处于发展阶段，如何在未知真实模型的结构和稀疏程度的情况下确定模型的结构和模型的稀疏程度，如何量化评价结构稀疏模型的好坏等问题，还需要长期的研究.

## 参 考 文 献


[1] Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society: Series B (Methodological), 2011, 73(3): 273-282

[2] Hosmer Jr D W, Lemeshow S. Applied Logistic Regression. USA New Jersey: John Wiley & Sons, 2004

[3] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67

[4] Engelhardt B E, Adams R P. Bayesian structured sparsity from Gaussian fields. arXiv preprint arXiv:1407.2235, 2014

[5] Ayaz U, Dirksen S, Rauhut H. Uniform recovery of fusion frame structured sparse signals. arXiv preprint arXiv:1407.7680, 2014

[6] Doshi-Velez F, Wallace B, Adams R. Graph-sparse LDA: A topic model with structured sparsity. arXiv preprint arXiv: 1410.4510, 2014

[7] Zhang S, Qian H, Zhang Z. A nonconvex approach for structured sparse learning. arXiv preprint arXiv: 1503.02164, 2015

[8] Chadwick J N, Bindel D S. An efficient solver for sparse linear systems based on rank-structured cholesky factorization. arXiv preprint arXiv:1507.05593, 2015

[9] Qiu C, Vaswani N, Lois B. Recursive robust PCA or recursive sparse recovery in large but structured noise. IEEE Transactions on Information Theory, 2014, 60(8): 5007-5039

[10] Oymak S, Jalali A, Fazel M. Simultaneously structured models with application to sparse and low-rank matrices. IEEE Transactions on Information Theory, 2015, 61(5): 2886-2908

[11] Wimalajeewa T, Eldar Y C, Varshney P K. Subspace recovery from structured union of subspaces. IEEE Transactions on Information Theory, 2015, 61(4): 2101-2114

[12] Soni A, Haupt J. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. IEEE Transactions on Information Theory, 2014, 60(1): 133-149

[13] Sun X, Qu Q, Nasrabadi N M. Structured priors for sparse-representation-based hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 2014, 11(7): 1235-1239

[14] Kyrillidis A, Baldassarre L, El Halabi M. Structured sparsity: Discrete and convex approaches//Boche H, Calderbank R, Kutyniok G, Vybíral J eds. Compressed Sensing and its Applications. Springer International Publishing, Switzerland: Lausanne, 2015: 341-387

[15] Vila J P. Empirical-Bayes Approaches to Recovery of Structured Sparse Signals via Approximate Message Passing [Ph. D. dissertation]. The Ohio State University, Ohio, USA, 2015

[16] Bai T, Li Y. Robust visual tracking using flexible structured sparse representation. IEEE Transactions on Industrial Informatics, 2014, 10(1): 538-547

[17] Hegde C, Indyk P, Schmidt L. A nearly-linear time framework for graph-structured sparsity//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France, 2015: 928-937

[18] Huang J, Zhang T. The benefit of group sparsity. The Annals of Statistics, 2010, 38(4): 1978-2004

[19] Sra S. Fast projections onto mixed-norm balls with applications. Data Mining and Knowledge Discovery, 2012, 25(2): 358-377

[20] Lazzaro D, Montefusco L B, Papi S. Blind cluster structured sparse signal recovery: A nonconvex approach. Signal Processing, 2015, 109(4): 212-225

[21] Vogt J, Roth V. A complete analysis of the $L_{1,p}$ Group-Lasso. arXiv preprint arXiv:1206.4632, 2012

[22] Rakotomamonjy A, Flamary R, Gasso G, Stephane C. $L_p-L_q$ penalty for sparse linear and sparse multiple kernel multi-task learning. IEEE Transactions on Neural Networks, 2011, 22(8): 1307-1320

[23] Vogt J E, Roth V. The group-Lasso: $l_{1,\infty}$ regularization versus $l_{1,2}$ regularization//Proceedings of the 32nd DAGM Conference on Pattern Recognition. Darmstadt, Germany: Springer-Verlag, 2010: 252-261

[24] Obozinski G, Jacob L, Vert J P. Group Lasso with overlaps: The latent Group Lasso approach. arXiv preprint arXiv: 1110.0413, 2011

[25] Jenatton R, Audibert J Y, Bach F. Structured variable selection with sparsity-inducing norms. The Journal of Machine Learning Research, 2011, 12(1): 2777-2824

[26] Percival D. Structured, sparse aggregation. Journal of the American Statistical Association, 2012, 107(498): 814-823

[27] Wang H, Leng C. A note on adaptive Group Lasso. Computational Statistics and Data Analysis, 2008, 52(12): 5277-5286

[28] Wei F, Huang J. Consistent group selection in high-dimensional linear regression. Bernoulli, 2010, 16(4): 1369-1384

[29] Meier L, van de Geer S, Bühlmann P. The Group Lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1): 53-71

[30] Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics, 2007, 23(12): 1486-1494

[31] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360

[32] Huang J, Ma S, Xie H, Zhang C H. A group bridge approach for variable selection. Biometrika, 2009, 96(2): 339-355

[33] Fu W J. Penalized regressions: The bridge versus the Lasso. Journal of Computational and Graphical Statistics, 1998, 7(3): 397-416

[34] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Statistical Science, 2012, 27(4): 481-499

[35] Zhang C H, Zhang S S. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society, 2014, 76(1): 217-242

[36] Huang J, Ma S, Xie H, Zhang C H. Amendments and corrections: A group bridge approach for variable selection. Biometrika, 2009, 96(4): 1024-1024

[37] Breheny P. The group exponential lasso for bi-level variable selection. Biometrics, 2015, 71(3): 1-10

[38] Friedman J, Hastie T, Tibshirani R. A note on the Group Lasso and a sparse Group Lasso. arXiv preprint arXiv: 1001.0736, 2010

[39] Yuan X T, Liu X, Yan S. Visual classification with multi-task joint sparse representation. IEEE Transactions on Image Processing, 2012, 21(10): 4349-4360

[40] Gao S, Chia L T, Tsang I W H. Multi-layer group sparse coding—For concurrent image classification and annotation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2011: 2809-2816

[41] Seetharaman I. Consistent Bi-Level Variable Selection via Composite Group Bridge Penalized Regression [Ph. D. dissertation]. Kansas State Univesity, Kansas, USA, 2013

[42] Liu M, Zhang D, Yap P T. Tree-guided sparse coding for brain disease classification//Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2012). Berlin Heidelberg, Germany: Springer, 2012: 239-247

[43] Liu J, Ye J. Moreau-Yosida regularization for grouped tree structure learning//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1459-1467

[44] Kim S, Xing E P. Tree-guided Group Lasso for multi-task regression with structured sparsity//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 543-550

[45] Turlach B A, Venables W N, Wright S J. Simultaneous variable selection. Technometrics, 2005, 47(3): 349-363

[46] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 2008, 9(3): 432-441

[47] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996, 381(6): 607-609

[48] Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research, 1997, 37(23): 3311-3325

[49] Olshausen B A, Field D J. How close are we to understanding V1? Neural Computation, 2005, 17(8): 1665-1699

[50] Simoncelli E P, Olshausen B A. Natural image statistics and neural representation. Annual Review of Neuroscience, 2001, 24(1): 1193-1216

[51] Jenatton R, Obozinski G, Bach F. Structured sparse principal component analysis//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010). Chia Laguna Resort, Sardinia, Italy: JMLR Proceedings, 2010: 366-373

[52] Ramirez I, Sprechmann P, Sapiro G. Classification and clustering via dictionary learning with structured incoherence and shared features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2010: 3501-3508

[53] Yu G, Sapiro G, Mallat S. Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity. IEEE Transactions on Image Processing, 2012, 21(5): 2481-2499

[54] Mairal J, Bach F, Ponce J, Sapiro G. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research, 2010, 11(1): 19-60

[55] Mairal J, Jenatton R, Obozinski G, Bach F. Network flow algorithms for structured sparsity//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1558-1566

[56] Mairal J, Jenatton R, Obozinski G, Bach F. Convex and network flow optimization for structured sparsity. Journal of Machine Learning Research, 2011, 12(9): 2681-2720

[57] Mairal J, Bach F, Ponce J. Task-driven dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 791-804

[58] Zhu S-C, Guo C-E, Wang Y, Xu Z. What are textons? International Journal of Computer Vision, 2005, 62(1-2): 121-143

[59] Zeiler M D, Krishnan D, Taylor G W. Deconvolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2010: 2528-2535

[60] Zeiler M D, Taylor G W, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning//Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011). Barcelona, Spain, 2011: 2018-2025

[61] Bo L, Ren X, Fox D. Multipath sparse coding using hierarchical matching pursuit//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2013: 660-667

[62] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2005: 886-893

[63] Ren X, Ramanan D. Histograms of sparse codes for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, 2013: 3246-3253

[64] Rigamonti R, Sironi A, Lepetit V, Fua P. Learning separable filters//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Jersey, USA, 2013: 2754-2761

[65] Kavukcuoglu K, Sermanet P, Boureau Y-L, et al. Learning convolutional feature hierarchies for visual recognition//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1090-1098

[66] Garrigues P, Olshausen B A. Group sparse coding with a Laplacian scale mixture prior//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 676-684

[67] Gregor K, Szlam A, LeCun Y. Structured sparse coding via lateral inhibition//Proceedings of the 28th International Conference on Machine Learning (ICML 2011). Bellevue, Washington, USA: Omnipress, 2011: 1116-1124

[68] Raman S, Fuchs T J, Wild P J, et al. The Bayesian group-Lasso for analyzing contingency tables//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 881-888

[69] Chandran M. Analysis of Bayesian Group-Lasso in Regression Models [Ph. D. dissertation]. University of Florida, Florida, USA, 2011

[70] Huang J, Zhang T, Metaxas D. Learning with structured sparsity. Journal of Machine Learning Research, 2011, 12(11): 3371-3412

[71] Bach F R. Structured sparsity-inducing norms through submodular functions//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 118-126

[72] Micchelli C A, Morales J M, Pontil M. Regularizers for structured sparsity. Advances in Computational Mathematics, 2013, 38(3): 455-489

[73] Bach F R. Consistency of the Group Lasso and multiple kernel learning. The Journal of Machine Learning Research, 2008, 9(1): 1179-1225

[74] Sharma D B, Bondell H D, Zhang H H. Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics, 2013, 22(2): 319-340

[75] Mairal J, Yu B. Supervised feature selection in graphs with path coding penalties and network flows. Journal of Machine Learning Research, 2013, 14(1): 2449-2485

[76] Rosset S, Zhu J. Piecewise linear regularized solution paths. The Annals of Statistics, 2007, 35(3): 1012-1030

[77] Lv X, Bi G, Wan C. The group lasso for stable recovery of block-sparse signal representations. IEEE Transactions on Signal Processing, 2011, 59(4): 1371-1382

[78] Jacob L, Obozinski G, Vert J P. Group Lasso with overlap and graph lasso//Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009). Montreal, Canada, 2009: 433-440

[79] Vincent M, Hansen N R. Sparse group lasso and high dimensional multinomial classification. Computational Statistics & Data Analysis, 2014, 71(3): 771-786

[80] Geng Z, Wang S, Yu M, et al. Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study. Biometrics, 2015, 71(1): 53-62

[81] Wu F, Yuan Y, Rui Y. Annotating web images using NOVA: NOn-conVex group spArsity//Proceedings of the 20th ACM Multimedia Conference (MM'12). Nara, Japan, 2012: 509-518

[82] Zeng L, Xie J. Group variable selection via SCAD-$L_2$. Statistics, 2014, 48(1): 49-66

[83] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94(1): 19-35

[84] Simon N, Friedman J, Hastie T. A sparse-Group Lasso. Journal of Computational and Graphical Statistics, 2013, 22(2): 231-245

[85] Zhou H, Sehl M, Sinsheimer J, Lange K. Association screening of common and rare genetic variants by penalized regression. Bioinformatics, 2010, 26(19): 2375-2382

[86] Sohn K A, Kim S. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012). La Palma, Canary Islands, 2012: 1081-1089

[87] Swirszcz G, Lozano A C. Multi-level lasso for sparse multi-task regression//Proceedings of the 29th International Conference on Machine Learning (ICML 2012). Edinburgh, UK, 2012: 361-368

[88] Wang J, Ye J. Two-layer feature reduction for Sparse-Group Lasso via decomposition of convex sets//Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. Montreal, Canada, 2014: 2132-2140

[89] Caner M, Han X. Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. Journal of Business & Economic Statistics, 2014, 32(3): 359-374

[90] Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing, 2015, 25(2): 173-187

[91] Chen X, Lin Q, Kim S, et al. Smoothing proximal gradient method for general structured sparse regression. The Annals of Applied Statistics, 2010, 6(2): 719-752

[92] Nesterov Y. Smooth minimization of non-smooth functions. Mathematical Programming, 2005, 103(1): 127-152

[93] Jenatton R, Mairal J, Obozinski G, Bach F. Proximal methods for hierarchical sparse coding. Journal of Machine Learning Research, 2011, 12(7): 2297-2334

[94] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202

[95] Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and Its Interface, 2009, 2(3): 369-380

[96] Tao Qing, Gao Qian-Kun, Jiang Ji-Yuan, Chu De-Jun. Survey of solving the optimization problems for sparse learning. Journal of Software, 2013, 24(11): 2498-2507(in Chinese)

(陶卿，高乾坤，姜纪远，储德军. 稀疏学习优化问题的求解综述. 软件学报，2013，24(11)：2498-2507)

[97] Rish I, Grabarnik G. Sparse Modeling: Theory, Algorithms, and Applications. New York, USA: CRC Press, 2014

[98] Rao N, Nowak R, Cox C. Classification with the sparse Group Lasso. IEEE Transactions on Signal Processing, 2016, 64(2): 448-463

[99] Wahab S, Zakaria M N, Sidek D, et al. Evaluation of auditory hallucinations in patients with schizophrenia: A validation study of the Malay version of Psychotic Symptom Rating Scales (PSYRATS). Psychiatry Research, 2015, 228(3): 462-467

[100] Huang S, Li J, Ye J, et al. A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6): 1328-1342

[101] Honorio J, Ortiz L, Samaras D, et al. Sparse and locally constant Gaussian graphical models//Proceedings of the Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Vancouver, Canada: Curran Associates, 2009: 745-753

[102] Honorio J, Samaras D, Rish I, Cecchi G. Variable selection for Gaussian graphical models//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012). La Palma, Canary Islands: JMLR Proceedings, 2012: 538-546

[103] Klami A, Virtanen S, Leppäaho E, et al. Group factor analysis. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(9): 2136-2147

[104] Chatterjee S, Steinhaeuser K, Banerjee A. Sparse Group Lasso: Consistency and climate applications//Proceedings of the 12th SIAM International Conference on Data Mining. Anaheim, USA, 2012: 47-58

[105] Gao L, Zhou S, Guan J. Effectively classifying short texts by structured sparse representation with dictionary filtering. Information Sciences, 2015, 323(11): 130-142

[106] Downey D, Bhagavatula C, Yang Y. Efficient methods for inferring large sparse topic hierarchies//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, 2015: 774-784

[107] Zheng M, Bu J, Chen C, et al. Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing, 2011, 20(5): 1327-1336

[108] Liu H, Yu H, Deng Z H. Multi-document summarization based on two-level sparse representation model//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA. 2015: 196-202

[109] Pnevmatikakis E A, Gao Y, Soudry D. A structured matrix factorization framework for large scale calcium imaging data analysis. arXiv preprint arXiv:1409.2903, 2014

[110] Han Y, Wu F, Tian Q. Image annotation by input—output structural grouping sparsity. IEEE Transactions on Image Processing, 2012, 21(6): 3066-3079

[111] Moayedi F, Azimifar Z, Boostani R. Structured sparse representation for human action recognition. Neurocomputing, 2015, 161(8): 38-46

[112] Ly N H, Du Q, Fowler J E. Sparse graph-based discriminant analysis for hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(7): 3872-3884

[113] Hegde C, Indyk P, Schmidt L. A fast approximation algorithm for tree-sparse recovery//Proceeding of the IEEE International Symposium on Information Theory (ISIT). New Jersey, USA, 2014: 1842-1846

[114] Bui H Q, La C N H, Do M N. A fast tree-based algorithm for compressed sensing with sparse-tree prior. Signal Processing, 2015, 108(3): 628-641

[115] Chen H, Wan Q, Fan R. Double-constraint flexible tree search-based orthogonal matching pursuit for DOA estimation using dynamic sensor arrays. International Journal of Electronics, 2015, 103(5): 1-9

[116] Castro R M, Tanczos E. Adaptive sensing for estimation of structured sparse signals. IEEE Transactions on Information Theory, 2015, 61(4): 2060-2080

[117] Peleg T, Eldar Y C, Elad M. Exploiting statistical dependencies in sparse representations for signal recovery. IEEE Transactions on Signal Processing, 2012, 60(5): 2286-2303

[118] Gui J, Tao D, Sun Z. Group sparse multiview patch alignment framework with view consistency for image classification. IEEE Transactions on Image Processing, 2014, 23(7): 3126-3137

[119] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning problems. Statistics and Computing, 2015, 25(6): 1129-1141

[120] Scheetz T E, Kim K Y A, Swiderski R E, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences, 2006, 103(39): 14429-14434

[121] Segal M R, Dahlquist K D, Conklin B R. Regression approaches for microarray data analysis. Journal of Computational Biology, 2003, 10(6): 961-980

[122] Sæbø S, Almøy T, Aarøe J, et al. ST-PLS: A multi-directional nearest shrunken centroid type classifier via PLS. Journal of Chemometrics, 2008, 22(1): 54-62

[123] Graham K, de Las Morenas A, Tripathi A, et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. British Journal of Cancer, 2010, 102(8): 1284-1293

[124] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 1999, 96(12): 6745-6750

[125] Singh D, Febbo P G, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 2002, 1(2): 203-209

[126] Gorman R P, Sejnowski T J. Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks, 1988, 1(1): 75-89

[127] Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. The Journal of Machine Learning Research, 2010, 11(3): 1081-1107

[128] Zhang T. Multi-stage convex relaxation for feature selection. arXiv preprint arXiv:1106.0565, 2011

[129] Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted $L_1$ minimization. Journal of Fourier Analysis and Applications, 2008, 14(5-6): 877-905

[130] Geman D, Yang C. Nonlinear image recovery with half-quadratic regularization. IEEE Transactions on Image Processing, 1995, 4(7): 932-946

[131] Foucart S, Lai M J. Sparsest solutions of underdetermined linear systems via $L_q$ minimization for $0<q\leq 1$. Applied and Computational Harmonic Analysis, 2009, 26(3): 395-407

[132] Yang H, Xu Z, King I. Online learning for Group Lasso// Proceedings of the 27th International Conference on Machine Learning. Scotland, UK: Omnipress, 2010: 1191-1198

[133] van de Geer S A, Bühlmann P. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 2009, 3(11): 1360-1392

[134] Zhang T. Some sharp performance bounds for least squares regression with $L_1$ regularization. The Annals of Statistics, 2009, 37(5): 2109-2144

[135] Ye F, Zhang C H. Rate minimaxity of the Lasso and dantzig selector for the $L_q$ loss in $L_r$ balls. The Journal of Machine Learning Research, 2010, 11: 3519-3540

[136] Chandrasekaran V, Parrilo P A, Willsky A S. Latent variable graphical model selection via convex optimization. The Annals of Statistics, 2012, 40(4): 1935-1967

[137] Ma S, Xue L, Zou H. Alternating direction methods for latent variable gaussian graphical model selection. Neural Computation, 2013, 25(8): 2172-2198

[138] Lauritzen S, Meinshausen N. Discussion: Latent variable graphical model selection via convex optimization. The Annals of Statistics, 2012, 40(4): 1973-1977

[139] Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 2010, 97(3): 519-538

[140] Fu F, Zhou Q. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. Journal of the American Statistical Association, 2013, 108(501): 288-300

[141] Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. The Annals of Statistics, 2010, 38(3): 1287-1319

[142] Wainwright M J, Ravikumar P, Lafferty J D. High-dimensional graphical model selection using $L_1$-regularized logistic regression//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006: 1465-1472

[143] Liu H, Han F, Yuan M. High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics, 2012, 40(4): 2293-2326

[144] Allen G I, Liu Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data// Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Philadelphia, USA, 2012: 1-6

[145] Yang E, Ravikumar P D, Allen G I. Graphical models via generalized linear models//Proceedings of the Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates, 2012: 1367-1375

[146] Liu Q, Ihler A T. Learning scale free networks by reweighted $\ell_1$ regularization//Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 40-48

[147] Danaher P, Wang P, Witten D M. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013, 76(2): 373-397

[148] Yang S, Pan Z, Shen X. Fused multiple graphical lasso. arXiv preprint arXiv:1209.2139, 2012

[149] Zhang B, Wang Y. Learning structural changes of Gaussian graphical models in controlled experiments. arXiv preprint arXiv:1203.3532, 2012

[150] Honorio J, Samaras D. Multi-task learning of Gaussian graphical models//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress, 2010: 447-454

[151] Committee on the Analysis of Massive Data et al. Frontiers in Massive Data Analysis. Washington, USA: National Academies Press, 2013

[152] Franke B, Plante J F, Roscher R, et al. Statistical Inference, Learning and Models in Big Data. arXiv preprint arXiv: 1509.02900, 2015

[153] Hammer B, He H, Martinetz T. Learning and modeling big data//Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium, 2014: 105-110

[154] Richtárik P, Takáč M. Distributed coordinate descent method for learning with big data. arXiv preprint arXiv: 1310.2059, 2013

[155] Richtárik P, Takáč M. Parallel coordinate descent methods for big data optimization. Mathematical Programming, 2016, 156(1): 433-484

[156] Mahajan D, Keerthi S S, Sundararajan S. A distributed block coordinate descent method for training $\ell_1$ regularized linear classifiers. arXiv preprint arXiv:1405.4544, 2014

[157] Kang D, Lim W, Shin K, et al. Data/feature distributed stochastic coordinate descent for logistic regression// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 1269-1278

[158] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122

**LIU Jian-Wei**, born in 1966, Ph.D., associate professor. His main research interests include machine learning, intelligent information processing, and analysis, prediction, controlling of complicated nonlinear system.

**CUI Li-Peng**, born in 1990, M.S. candidate. His main research interest is machine learning.

**LUO Xiong-Lin**, born in 1963, Ph.D., professor. His main research interests include intelligent control, and analysis, prediction, controlling of complicated nonlinear system.

**Background**

Nowadays, High-dimensional and small sample datasets, where the number of unknown variables which are to be estimated as one or several orders of magnitude larger than the number of samples in the data, are the rule rather than the exception in bioinformatics, psychology diagnosis, computational linguistics and phonetics, computer vision, the Portal site, e-commerce, mobile Internet, and Internet of things. Classical statistical estimation and inference cannot be used for high-dimensional and ultra-dimensional problems. By introducing additional structural smoothness assumptions, or say restricting to a certain models class of smooth functions, Lasso method is well-established framework for fitting of a linear model having many more unknown parameters than observations.

The Lasso ignores the structure of the variables and can only achieve sparsity on the level of variables. By inducing structured sparsity or low rank or those based on more general loss functions, the structured sparsity models considers the structure of the variables as the prior information and often achieves better statistic properties than the regular sparse models.

This survey brings together methodological concepts, computational algorithms, applications and mathematical theory for structured sparsity model of high-dimensional datasets. We give a systematical survey of the structure sparse models and it's solving algorithms. We point out different structure sparse models, advantages and disadvantages and their specific application scenes. We also show the conditions which are often used when discussing the statistic properties such as variable selection consistency and parameter estimation consistency. Finally, we give the meaningful directions of the future research in the field of structured sparsity learning.

This work is supported by the Basic Scientific Research Foundation of China University of Petroleum (JCXK-2011-07).